AI Engineering

Progressing towards Robust and Trustworthy AI Systems

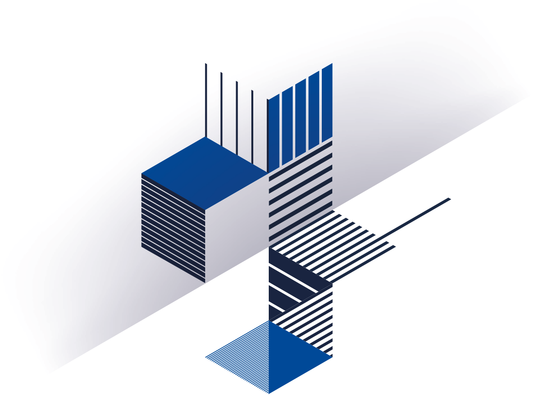
ANNUAL REPORT 2020

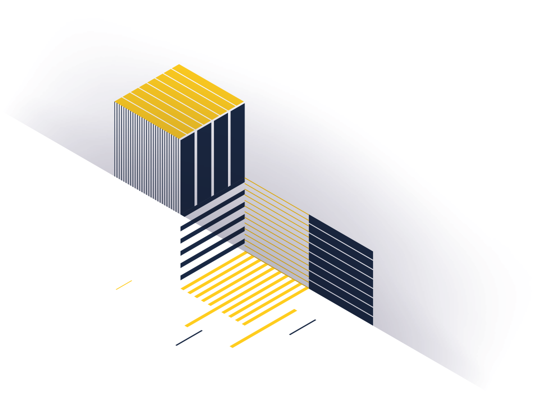
fortiss

> I am an optimist and I believe that we can create AI for the good of the world. That it can work in harmony with us. We simply need to be aware of the dangers, identify them, employ the best possible practice and management, and prepare for its consequences well in advance.

Stephen Hawkins at Web Summit in Nov. 2017

TABLE OF CONTENTS





1. INTRODUCTION 6

Software (including AI) Eats the World 7

Need for AI Engineering 7

Challenges of AI Engineering. 8

AI Engineering 9

Benefits of AI Engineering 9

Contributions to AI Engineering 10

Acknowledgements. 12

2. RESEARCH LINES 14

2.1 Structured Approach for Trustworthy AI
Dr. Henrik Putzer, Dr. Ernest Wozniak 15

2.2 Knowledge-Augmented Machine Learning
Dr. Julian Wörmann, Dr. habil. Hao Shen, Alexander Sagel, Amit Sahu 25

2.3 Joint Action Planning
Klemens Esterle, Patrick Hart, Tobias Kessler 30

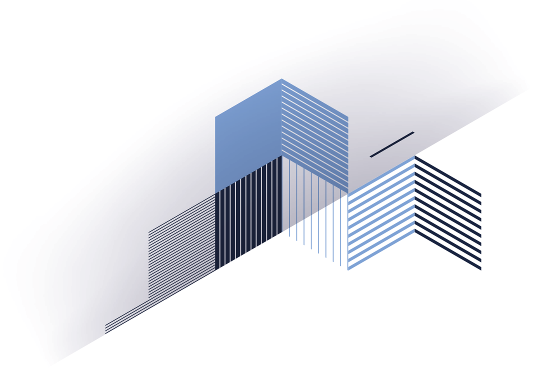
2.4 Verification of Machine Intelligence
Dr. Mojdeh Golagha, Amit Sahu 35

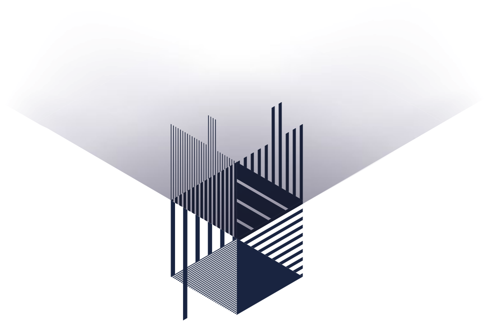
2.5 Human-centered Machine Learning
Dr. Yuanting Liu, Dr. habil. Hao Shen. 43

2.6 Automated Program Synthesis
Dr. Harald Rueß, Dr. Xin Ye. 47

2.7 Edge AI
Prof. Dr. Rute Sofia. 56

2.8 Neuromorphic Computing
Axel von Arnim, Manos Angelidis 63





3. APPLICATIONS 72

3.1 fortiss Labs
Dr. Markus Duchon 73

3.1.1 Industrial IoT Lab
Prof. Dr. Rute Sofia 74

3.1.2 Robotics Lab
Alexander Perzylo, Dr. Markus Rickert 78

3.1.3 Energy Lab
Dr. Denis Bytschkow, Dr. Markus Duchon 81

3.1.4 Mobility Lab
Simon Barner, Tobias Kessler 86

3.1.5 Drone Lab
Dr. Ernest Wozniak, Florian Grötzner 90

3.2 IBM fortiss Center for AI
Dr. Holger Pfeifer 93

3.2.1 Accountable Federated Machine Learning
Dian Balta, Dr. habil. Ulrich Schöpp, Mahdi Sellami . . 93

3.2.2 Proactive & Interactionless Government Services
Dian Balta, Peter Kuhn 98

3.2.3 Human-centered Machine Learning
Dr. Yuanting Liu, Dr. habil Hao Shen, Sören Klingner, Zhiwei Han, Stefan Matthes, Tianming Qiu 102

3.2.4 Anomaly Detection in Robot-based Manufacturing with Semantic Digital Twins
Dr. Markus Rickert, Alexander Perzylo 106

4. TRANSFER 110

4.1 fortiss Mittelstand
Dr. Wolfgang Köhler 111

4.2 Center for Code Excellence
Dr. Johannes Kross 114

IMPRINT 116

1

INTRODUCTION

INTRODUCTION

Software (including AI) Eats the World

Software-intensive systems play an important role in many areas of modern life, including transportation, manufacturing, aerospace and healthcare. Indeed, innovations are increasingly based on high-performance code, the share of value creation achieved by software in embedded control systems is rising steadily, and societal-scale service and utility infrastructures are increasingly run by automated control software.

It is only in conjunction with AI-based software capabilities, however, that many applications develop their full potential. In particular, AI-based software systems may interpret perception stimuli, determine relevant information, build faithful environment models, select and prioritize the most appropriate goals, plan to achieve selected goals in an efficient manner, adapt behavior, goals and planning through learning and reasoning, and distill knowledge from experience. As a result, an AI system can not only perceive the physical environment but also have the ability to learn from the wealth of experience gained, to derive new insights, to understand contexts and to make important decisions in a supportive and increasingly autonomous manner.

These types of cognitive capabilities can be realized using a wide variety of technologies from the broad field of AI. In particular, techniques for searching and planning, optimization, logical and inductive reasoning, and approximation or interpolation are used for this purpose. Artificial neural network structures, in particular, are the currently dominating example-based program synthesis techniques for approximating ("learning") functions from pre-selected input-output grid points.

Need for AI Engineering

Armed with a wealth of AI techniques we are, by and large, able to build many impressive AI systems, including surgical robots, swarm-intelligent drone systems, or decentralized controls for smart service infrastructures (e.g., energy/ water). Clearly, building these AI system is more than just running a machine learning and/or planning algorithm, and building and running them is a serious undertaking.[1]

Despite technological advances that have led to the proliferation of AI-based solutions, questions remain about the level of trust that can be placed in AI systems. What is missing, therefore, is a rigorous approach to building and operating AI systems in which people can trust.

Developing such trustworthy AI, however, is difficult, expensive, and error-prone. A key reason for this difficulty is that AI systems

- continuously learn, adapt and optimize themselves based on experience,
- operate partially in unknown or uncertain environments,
- increasingly lack a fallback to a responsible (human) operator,

[1] Sculley et al, Hidden technical debt in ML systems, 2015

- offer a variety of new attack surfaces (e.g. sensor spoofing), and
- often lead to largely unpredictable and emergent behavior in operation.

In particular, increasingly autonomous AI systems are expected to robustly operate in the presence of inaccuracies, uncertainties, and errors in the model (uncertain knowledge) as well as in the presence of non-modeled phenomena (uncertain ignorance).[2]

Challenges of AI Engineering

A key difficulty is that we are currently lacking established methods, processes and tools for engineering trustworthy AI systems.

→ How can we distinguish between desired and undesired behavior of data-driven AI systems?

→ How can we design AI systems which operate robustly in uncertain or unknown environments?

→ How can we ensure accountability in distributed AI systems?

→ How can we safely compose AI systems from individual components, some of which may be learning-enabled?

→ How can we design architectures of AI systems with a meaningful integration of behavioral ("fast") and deliberative ("slow") cognitive capabilities?

→ How can we integrate knowledge into learning-enabled systems in a meaningful and useful manner?

→ How can we specify and verify AI components and systems?

→ What are the relevant data for specifying AI systems?

→ What kind of assurance is needed for AI systems? How to compile assurance cases in a compositional manner?

→ What are suitable metrics for measuring the intended performance of AI systems?

→ How can human operators interpret and interact with an AI system in a meaningful way?

→ How can we build performant AI systems for embedded and IoT systems applications?

→ How can we ensure meaningful (human) control over an AI system in operation?

[2] Infamously, the former US Secretary of Defense Donald Rumsfeld refered to these notions as the "known unknown" and the "unknown unknown", respectively.

→ How can we continuously, throughout the entire life cycle, ensure the safety and security of increasingly autonomous AI systems?

These kinds of fundamental challenges form the basis of AI Engineering. This emerging field of science has the goal of developing rigorous methods, processes and tools for engineering trustworthy AI systems, thereby facilitating their successful adoption in daily life.

AI Engineering

AI Engineering heavily builds on established methods of software, systems and dependability engineering. A naïve adoption of these established techniques for building software-intensive systems to AI systems, however, is often not possible. Code coverage metrics for example, can be established for an artificial neural network already using a single test case, and the real benefit of MC/DC coverage on the level of ReLu nodes in these networks is, at least, questionable. Moreover, current regimes for safety and security certification are not directly applicable to AI systems, since the approval of mission-critical systems requires that the system behavior, together with its operating environment, be fully specified and verified prior to commissioning.

Over the last five years,[3] we have been developing at fortiss rigorous engineering principles for enabling robust and trustworthy AI applications which

- safely operate in uncertain, unpredictable environments,
- make timely and reliable decisions whose results are comprehensible and explainable,
- is resilient to erroneous inputs and targeted attacks,
- processes ever-increasing amounts of data,
- but can also extract useful insights from small amounts of data, without significant compromises in confidentiality and privacy.

Benefits of AI Engineering

AI Engineering is crucial in expanding the range of current applications of AI technologies. In particular, sound AI engineering technology is a prerequisite for

- AI-based embedded software systems applications including cloud-based control of manufacturing processes and machinery, swarms of service drones, control of critical service and utilities infrastructures, ecosystem of learning autonomous vehicles, or surgical robots and
- AI-assisted decision support systems with transparent and meaningful human-AI interactions for automated risk management and root cause analysis in many novel applications including AI-based co-pilots, predictive maintenance or attack analysis and management.

AI Engineering however, does not only open completely new application areas for AI systems. We are also expecting significant benefits and break-through AI applications through

[3] See also Whitepaper, "Künstliche Intelligenz – Chancen und Risiken für Bayern", 2017 (https://www.fortiss.org/fileadmin/user_upload/Veroeffentlichungen/Informationsmaterialien/191029_fortiss_KI_White_Paper_web.pdf)

- Increased acceptance of AI-based solutions by means of comprehensible and explainable decisions, for example, for verifiable diagnostics in the medical field, comprehensible pricing in, say, energy auctions, secured s cenario-based decision support in self-operated transportation vehicles,
- A competitive advantage through further congruence with GDPR in personal data processing (e.g. automated lending),
- More efficient development and secure operation of AI solutions through cross-domain reuse of configurable AI building blocks and
- Supporting market access of mission-critical AI systems through a rigorous AI engineering methodology for continuous assurance and certification.

Contributions to AI Engineering

We have been structuring our program on AI Engineering along the fortiss triad of research, application, and transfer.[4]

Research

We have been developing eight interconnected and solution-oriented research lines for developing novel solutions to specific AI Engineering challenges towards robust and trustworthy AI.

A Structured Approach for Trustworthy AI

B Knowledge-Augmented Machine Learning

C Joint Action Planning

D Verification of Machine Intelligence

E Human-centered Machine Learning

F Automated Program Synthesis

G Edge AI

H Neuromorphic Computing

[4] In German: forschen. anwenden. gestalten.

Application

a. The results as developed in these research lines are tested and prototyped using the infrastructure and demonstrators of the fortiss labs. We are currently developing in close cooperation with industrial and academic partners, specific application-oriented labs for Industrial Internet of Things applications, intuitive programming of robots, smart control of energy systems, and self-driving cars.

b. The experiments and prototyping based on the software and hardware infrastructure of the fortiss labs are useful research drivers in that they allow us to demonstrate the applicability of research results, point at necessary improvements, and also trigger, sometimes rather unexpectedly, new research directions and challenges.

c. In cooperation with our partner IBM we are developing a portfolio of solutions-oriented AI projects to sustainably tap the potential of AI. The IBM fortiss Center for AI is globally networked with research and application partners from, among others, Germany, Switzerland, Ireland, US, and is currently addresses the following topics:

    i. Prototyping of hyperledger-based architectures for enabling accountability in federated Machine Learning,

    ii. Application of human-centered machine learning technology to stress management,

    iii. Neuro-symbolic integration for anomaly detection in robot-based manufacturing,

    iv. Distributed ledger technology for eGovernment applications.

Transfer

a. We are compiling best practices on AI engineering, and contribute to corresponding standardization efforts. In particular, fortiss has been one of the main contributors of the six volume engineering reference model VDE-AR-E 2842 for trustworthy AI, which is mainly based on the structured AI engineering approach developed at fortiss. Thereby, we are ensuring the widespread availability and world-wide applicability of state-of-the-art knowledge on AI Engineering.

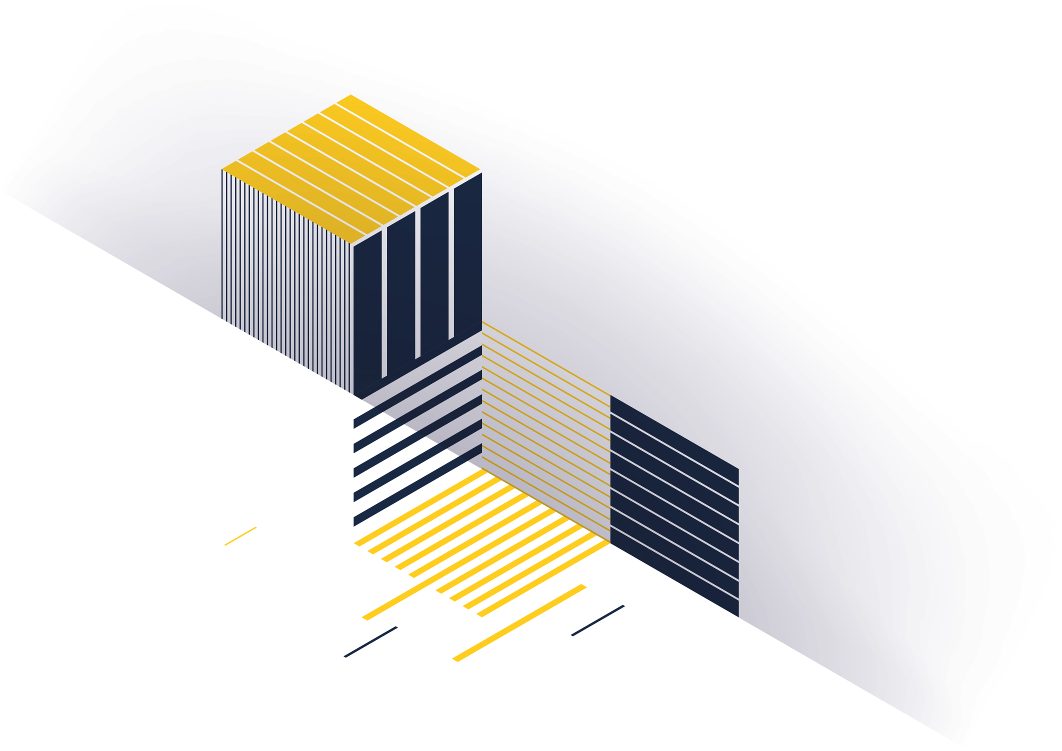
b. fortiss transfer centers offer a portfolio of information, qualification, and prototyping formats based on state-of-the-art findings and experience on AI Engineering. Thereby, the activities of the fortiss Mittelstand target software- and AI-related small and medium businesses, whereas the fortiss Center for Code Excellence teaches basic principles of AI engineering and coaches startup teams in the context of the TUM Venture Labs.

Here we are reporting on our progress and results on AI Engineering activities.

Acknowledgements

The fortiss AI research, application, and transfer program is funded by the Bavarian Ministry of Economics within the framework of the Free State of Bavaria digitization strategy Digital II.

> As a result, an AI system can not only
> perceive the physical environment
> but also have the ability to learn from the wealth
> of experience gained, to derive new insights,
> to understand contexts and to make important
> decisions in a supportive and increasingly
> autonomous manner.



2

RESEARCH LINES

RESEARCH LINES

2.1 Structured Approach for Trustworthy AI

Authors:
Dr. Henrik Putzer, Dr. Ernest Wozniak

Despite technological advances that have led to a proliferation of AI-based solutions, questions remain about the level of trust that can be placed in such software systems. What is missing, in particular, is a rigorous and structured approach to build and operate AI systems in which people can trust. In particular, attributes of trustworthiness including functional safety, cybersecurity, privacy, usability and maintainability as well as legal and ethical aspects are relevant.

In traditional safety engineering, the basis for certification is an assurance case. This is a structured and convincing argument for the trustworthiness and safety of the system under consideration—with respect to a well-defined operating environment, for pre-defined use cases and for the intended purpose and benefit. Such a structured argument with its evidences needs to be derived from a structured development approach delivering (in a structured, documented and reproducible manner) the system under consideration together with suitable development artefacts (e.g. design reviews, test reports). For technologies such as electronics and software the structured development approaches (processes, methods, artefacts) are prescribed through industrial standards such as the domain-agnostic IEC 615081.

Currently there is no such structured approach for developing technical systems based on AI. There is no generally accepted and documented development approach nor is there a generally accepted and documented way to ensure trustworthiness when it comes to the development of AI-based systems. Moreover, a relatively complete and continuous set of generally accepted methods and tools for supporting the complete life cycle for engineering AI-based systems is still lacking.

The goal of our Structured Approach for Trustworthy AI is to pull together the available expertise, research and experiences to define such a structured and generally accepted approach to trustworthy AI. With such a document, the industry would have guidance on how to develop and argument trustworthy AI application, and even more relevant, the predictability of legal decisions would increase. Furthermore, such a document could serve as a reference to define maturity models and acceptance criteria for trustworthy AI. In the end this could result in an explicit certificate for trustworthy AI products. After all, it is already difficult for end users to recognize the quality of products using conventional technologies. This is even harder when it comes to AI technologies. With a certificate on the trustworthiness of AI products end users could be guided on their buying decisions to pick the AI product with the level of trustworthiness needed. Overall the confidence in AI products could be increased.

Our approach for defining a structured approach to trustworthy AI is based on two main insights.

1. AI = new technology & new engineering approach. Even if AI in some cases appears to be magic, in science and engineering we do not deal with magic. AI is "just" a new engineering approach, a new technology with new characteristics (that are not even fully understood by AI scientists) and its specific development approach (e.g. engineering method) that produces good old automation with all the known problems as described by Billings 2.

2. AI is (only) one element within a system or product. AI never is a system on its own. It is always a component within the context of a system which in turn is situated within the context of an environment that needs to be considered. Hence, an adopted systems-engineering process needs to be considered for the development of AI based systems.

Consequently, a new structured approach (Putzer et al, 2020a; Putzer et al, 2020b; Putzer, 2019; Putzer, 2020a; Putzer, 2020b) needs to be based on modern systems-engineering and needs to consider trustworthiness aspects and new technologies to develop trustworthy autonomous/cognitive systems (A/C-systems) that are based on AI. To accomplish this, we use the ICE 61508 as a starting point. This international standard is the industry independent and generic master of all standards handling functional safety in electric, electronic and programmable systems (E/E/EPS systems).

Basically the IEC 61508 defines a risk-based approach along a reference life cycle with a structured approach (process) including requirements on measures and methods. This is the approach adopted and extended by our structured approach. The most relevant extensions and new key concepts are discussed in the next section which lead to the VDE-AR-E 2842-61 Development and Trustworthiness of Autonomous/Cognitive Systems (refer to VDE standard—VDE, 2020). This is a recently released document on trustworthy AI developed by the Association for Electrical, Electronic and Information Technologies (VDE) under the direction and sustainable contribution of fortiss.

New key concepts and extensions

The backbone of the VDE-AR-E 2842-61 is the risk-based approach along the trustworthiness reference life cycle (see Figure 1), which is derived from the life cycle of the IEC 61508. The main extensions and used key concepts can be described as follows:

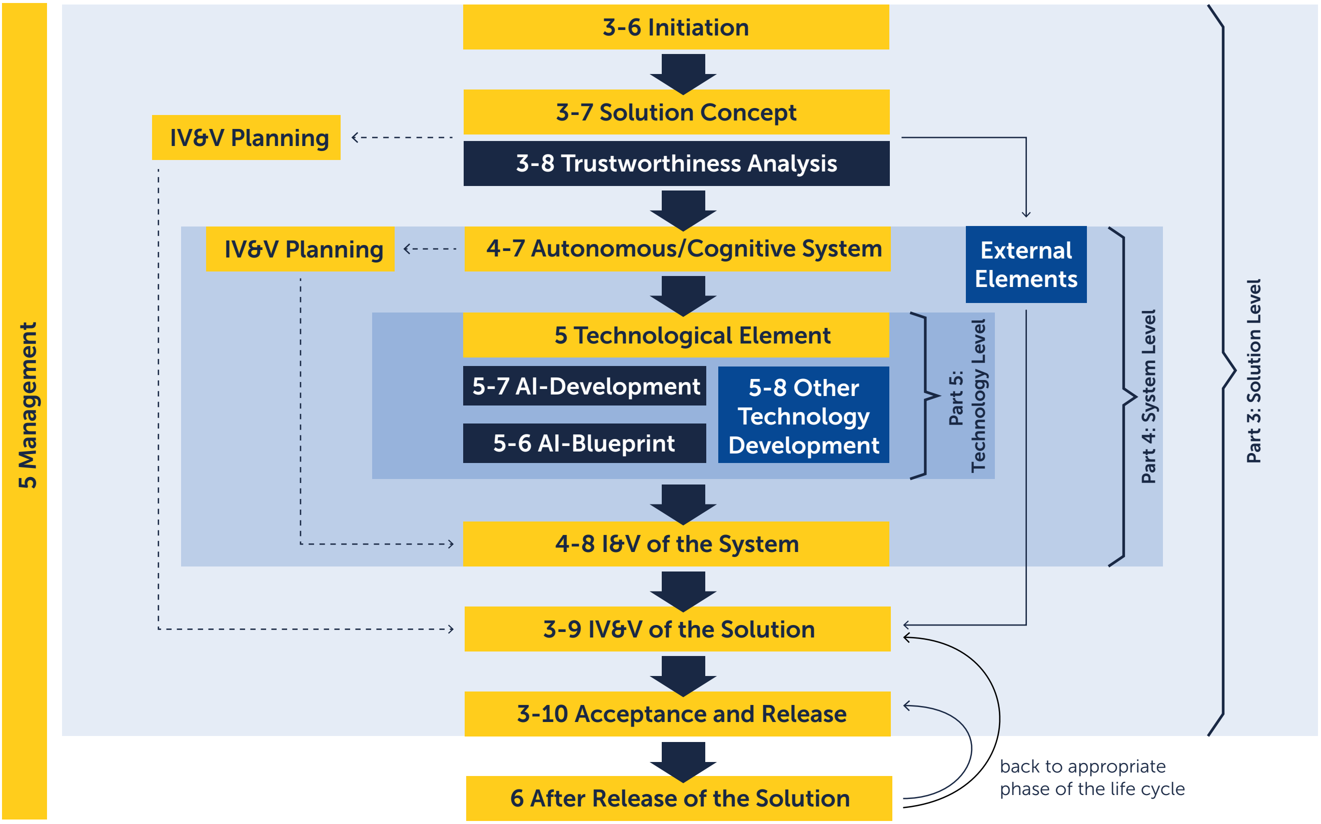


Figure 1. reference life cycle

From safety to trustworthiness

The IEC 61508 deals with functional safety. Currently a lot of aspects have to be handled during modern system development, such as safety (incl. functional safety and safety of the intended function), security, privacy, usability and ethics. In the VDE-AR-E 2842-61 this leads to the meta term trustworthiness (refer to VDE standard part 1—VDE, 2020) which is the per-project suitable selection and combination of aspects as indicated in Figure 2.

Solution level

The solution level (refer to VDE standard part 3—VDE, 2020) adds one level of abstraction above the A/C-system (see Figure 1, second block from top) which results in an additional phase in the trustworthiness reference life cycle. The solution includes the AI-system as a black box and examines its role and behavior in the overall environment including the user, other interface partners and

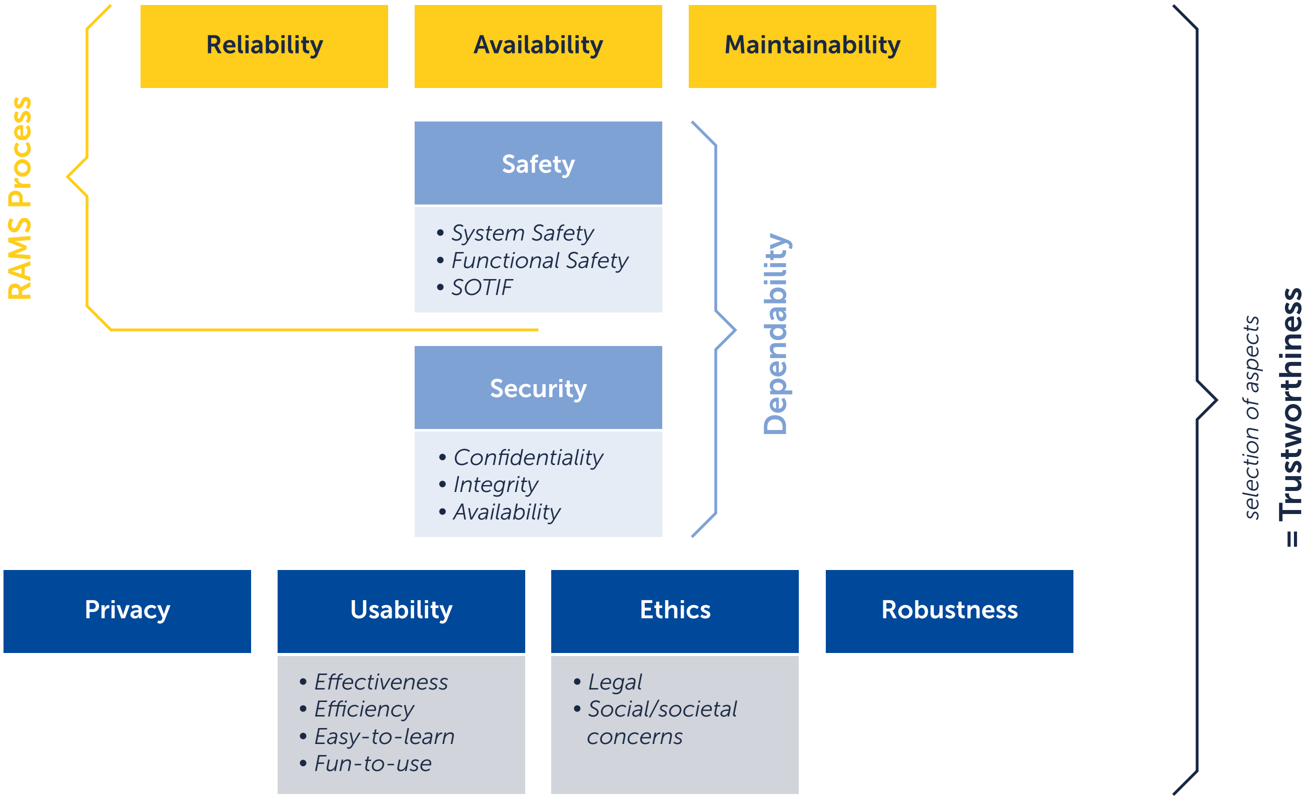


Figure 2.
Defining trustworthiness as a meta term of aspects

stakeholders. The solution level contains the ideas on the socio-technical work system[3] including all modelling and analyses.

This solution level is the origin of all hazards[4]. To identify and quantify all hazards, well known analyses of AI trustworthiness aspects are executed (e.g. hazard analysis and risk assessment—HARA or thread analysis and risk assessment—TARA) resulting in a list of hazards with heterogeneous attributes (e.g. fault tolerant time interval, safe state, SIL). These hazards are mitigated by a trustworthiness solution concept with specific mitigating measures.

System level

This design phase (refer to VDE standard part 4—VDE, 2020) defines a modern systems engineering approach. All defined activities and requirements are iteratively applied to form the hierarchical development of the A/C system. Each iteration—from A/C system via subsystem, component, sub-component etc.—refines the requirements and architecture (Cheng et al, 2019a) of the design and carefully keeps track of the trustworthiness measures and detailed trustworthiness functions and requirements and their trustworthiness attributes.

Technology level

This phase covers the development of an element using a specific (AI) technology. The main contribution of this phase is the concept of the *AI blueprint* (Wozniak et al, 2021). An AI blueprint is a generic and structured approach for the development of an AI element based on a certain AI technology (e.g. deep neuronal networks) motivated by the fact that AI is an additional type of technology that cannot be handled by existing development approaches.

An example for such an AI blueprint scoping the development of deep neuronal networks is provided in Figure 3. It includes a clear development contract including trustworthiness related assumptions and guarantees to ensure a seamless plug-in integration into the overall trustworthiness related development process. To ensure the quality of the AI blueprint and the trustworthiness of generated AI elements, the structured approach contains strict requirements on how to define and qualify an AI blueprint to even be able to cover any future new technology.

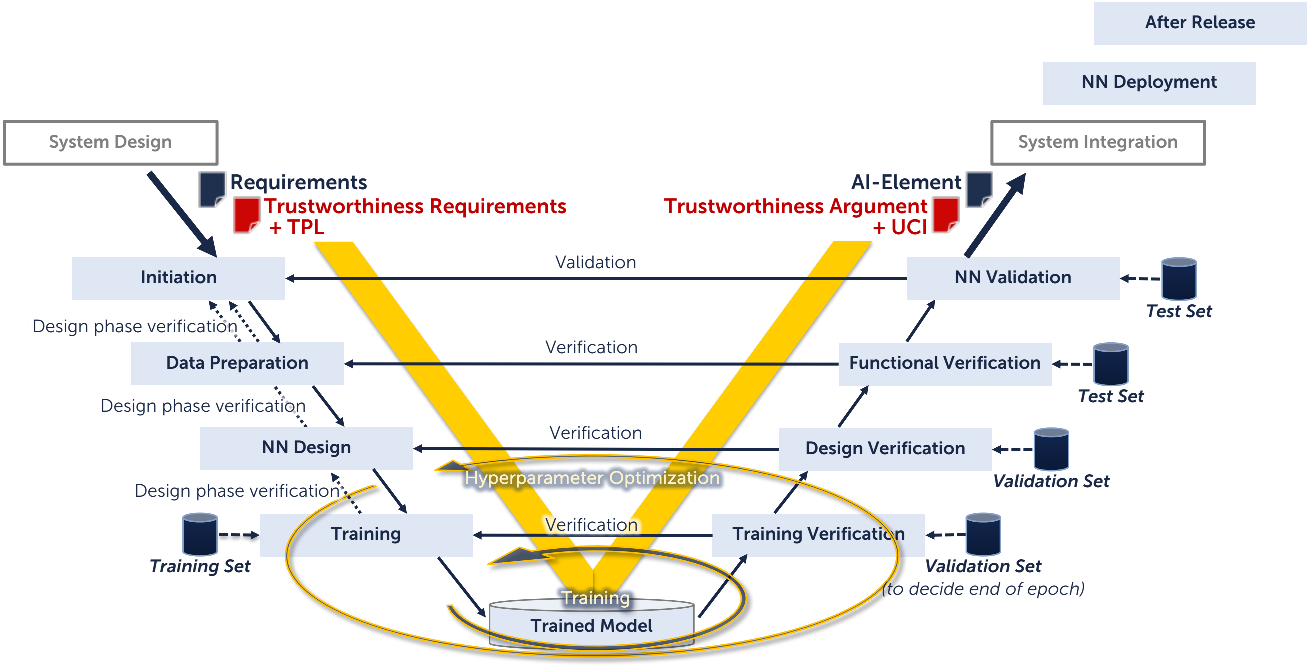


Figure 3.
Trustworthiness reference life cycle

Uncertainty confidence indicator (UCI)

Each type of technology has its own types and causes of failures. Current standards like the IEC 61508 propose that software only has systematic failures. Measures to avoid such systematic failures include a good development culture (e.g. safety culture), relying on experts and well-known designs, methods and measures ideally defined in a documented process. For electronic elements we see random failures as another cause. Quantitative measures (e.g. based on fault rates and fault tree analyses) and metrics like safe failure fraction or the diagnostic coverage help to develop safe (or even trustworthy) designs.

With some AI technologies (e.g. neuronal networks) we see a third type of failure: the *uncertainty-related failure*. This failure cannot be mitigated by good processes and established metrics. It is a new and characteristics-based type of failure that is inherent to the technology of neuronal networks and some other machine learning approaches (refer to VDE standard part 5—VDE, 2020). To handle this third kind of failure, the uncertainty confidence indicator (UCI) is introduced (see Figure 3 and VDE standard part 5—VDE, 2020).

Trustworthiness assurance case

Finally, the structured approach in the VDE-AR-E 2842-61 proposes the trustworthiness assurance case (refer to VDE standard part 3—VDE, 2020). Based on scientific research (Wozniak et al, 2020) this assurance case considers all trustworthiness aspects, the risk-based approach and the overall sound argumentation that the AI system is trustworthy in the defined use cases and environments. This argumentation is based on evidences that are derived from development artifacts generated during the development process (e.g. design verification reports, test reports).

type of failure	measures	HW measures	SW measures	AI measures
systematics	qualitative requirements	systematic capability	systematic capability	
random	quantitative requirements	λ, SFF, DC, target values	– / –	– / –
uncertainty-related	structured approach	– / –	– / –	uncertainty confidence indicator (UCI)

Figure 4.
Three types of failures and their mitigating measures

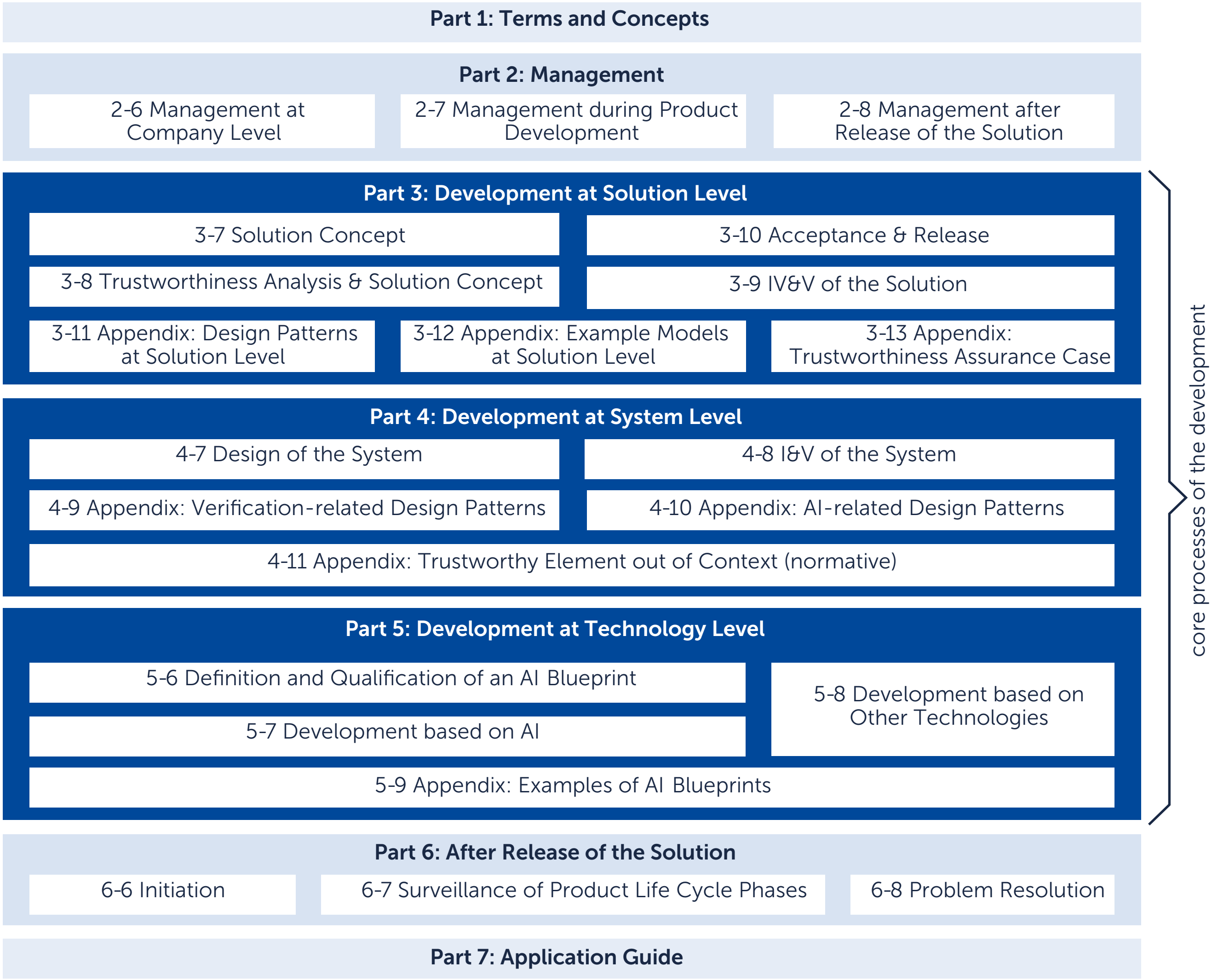
Results and benefits

The VDE-AR-E 2842-61 describes and determines the state-of-the-art in structured development of trustworthy AI-based systems. Consequently, this application rules provides the framework for developing AI-based products with a clear perspective of getting them certified for the market. The VDE-AR-E 2842-61 answers the question of how to build AI, how to verify AI (incl. trustworthiness assurance case) for developers, and provides a reference framework for how to certify AI-based systems. It clearly separates ethical fundamentals and societal acceptability from the technical approach. The VDE-AR-E 2842-61 provides a generic framework for the development of trustworthy solutions and trustworthy autonomous/cognitive systems. It defines a reference life cycle analagous to the key functional safety standards (i.e. IEC 61508) as a unified approach to achieve and maintain the overall performance of the solution and the intended behavior and trustworthiness of the autonomous/cognitive system.

The VDE-AR-E 2842-61 in its first version is about to be finalized. An overview of the standard is provided in Figure 5. The standard consists of seven parts. Parts 1, 2, 3 and 6 have been finalized and accepted by the working group. Parts 4 and 5 are scheduled to be finalized and accepted during the first quarter of 2021.

First scientific applications of the VDE-AR-E 2842-61 as in Putzer et al, 2019, delivered both: We got promising results in structuring the development of autonomous/cognitive systems, providing a framework to contextualize modern systems engineering approaches and AI related methods and measures. On the other side many questions arose concerning details in process interfaces, methods and application practice (see next section VI Conclusion and Future Work).

Figure 5.
Three types of failures and their mitigating measures



Part 1: Terms and Concepts

Part 2: Management

2-6 Management at Company Level

2-7 Management during Product Development

2-8 Management after Release of the Solution

Part 3: Development at Solution Level

3-7 Solution Concept

3-10 Acceptance & Release

3-8 Trustworthiness Analysis & Solution Concept

3-9 IV&V of the Solution

3-11 Appendix: Design Patterns at Solution Level

3-12 Appendix: Example Models at Solution Level

3-13 Appendix: Trustworthiness Assurance Case

Part 4: Development at System Level

4-7 Design of the System

4-8 I&V of the System

4-9 Appendix: Verification-related Design Patterns

4-10 Appendix: AI-related Design Patterns

4-11 Appendix: Trustworthy Element out of Context (normative)

Part 5: Development at Technology Level

5-6 Definition and Qualification of an AI Blueprint

5-8 Development based on Other Technologies

5-7 Development based on AI

5-9 Appendix: Examples of AI Blueprints

Part 6: After Release of the Solution

6-6 Initiation

6-7 Surveillance of Product Life Cycle Phases

6-8 Problem Resolution

Part 7: Application Guide

The VDE-AR-E 2842-61 has been applied in parts. We were able to structure several projects and add the benefits of the solution level in higher level requirements analysis and risk analyses. Furthermore, the entries processes of a company for the development of an engineering process that is driven by data could be improved and structured. Furthermore, there are initial approaches for using the VDE-AR-E 2842-61 as a reference to generate checklists targeting trustworthiness assessments for existing designs.

Ongoing development

Although the VDE-AR-E 2842-61 is available as a stable and rigid framework for the development of A/C systems, there are some topics for future evolutions of the standard.

apply ~ eval ~ improve: Now that the VDE-AR-E 2842-61 is available it should find widespread application in a large number of industrial projects from various application domains. Further experience with applying this development framework yields new insights which are expected to be included in further editions of the VDE-AR-E 2842-61 application rules. In particular, further experience in developing these kinds of systems will lead to more detailed tool boxes for metrics, development methods and verification measures. However, in updating this application rule we still need to ensure that the standard remains applicable with small overhead so that small and medium companies (SMEs) are also able to apply it to their product development process.

Research on AI

The content of the VDE-AR-E 2842-61 needs to be refined. A lot more detail is necessary to improve the measures, methods and metrics on AI technologies. And more breadth is necessary, e.g. more AI blueprints to cover more AI technologies beyond the content (refer to VDE standard part 5—VDE, 2020). Further research is needed to better understand AI technologies such as neuronal networks and to understand how trustworthiness can be measured (UCI), ensured and included in an assurance case. In addition, further topics such as continuous integration, high number of variants in products and reusability need to be supported.

Integrate knowledge from other working groups

VDE/DKE are the first organizations to come up with a standard on structured development of trustworthy AI-based systems, but certainly they are not the only one. Many groups are working on this topic including ISO/IEC JTC1 SC42 (see *Deutsche Normungsroadmap Künstliche Intelligenz*[5]). It would seem that synchronizing the concepts and solutions of these initiatives into a single, and hopefully harmonized and structured approach to trustworthy AI, would be a worthwhile effort.

[5] Wahlster W., Winterhalter C. (Herausgeber): Deutsche Normungsroadmap Künstliche Intelligenz, DIN & DKE, 2020-11

Internationalization

The VDE-AR-E 2842-61 is a national standard. The goal however, is to internationalize the standard through ISO or IEC. Apart from the activities to integrate and harmonize the knowledge from other working groups, there are currently activities by, by Japan and others, to adopt the VDE-AR-E 2842-61 as their respective national standard on trustworthy AI as well.

Certification

When it comes to certification and homologation of products, the VDE-AR-E 2842-61 serves as a reference model. Questions can be answered such as what is the certification interface between developers and certification and how to validate AI based systems with safety-relevance or trustworthiness requirements. In the long run specific AI related certificates can be developed. Such certificates help less experienced people and users to make buying decisions and to establish trust in AI based products. Overall these certificates should enhance acceptance of and confidence in AI-based products, thereby leveraging the economic success of AI-based products and services.

Call for participation

Finally, if you and/or your organization are interested in discussing any topic related to the VDE-AR-E 2842-61, in contributing to the evolution of it, in applying this standard to your development project, or in pushing the state-of-the-art in autonomous/cognitive systems engineering further, please contact us. We would be more than happy to support you.

LITERATURE

Cheng, C., Gulati, D. , & Yan, R. (2019a). Architecting Dependable Learning-enabled Autonomous Systems: A Survey. arXive, ID 1902.10590, 2019.

Putzer, H.J. & Wozniak, E. (2020a). A Structured Approach to Trustworthy Autonomous/Cognitive Systems. arXiv preprint arXiv:2002.08210.

Putzer, H.J. & Wozniak, E. (2020b). Trustworthy Autonomous/Cognitive Systems—A Structured Approach, white paper, https://www.fortiss.org/veroeffentlichungen/whitepaper, 2020-10.

Putzer, H.J. (2019). Ein strukturierter Ansatz für verlässliche KI, Key-Note, VDE-DKE-Tagung Funktionale Sicheheit für die Zulunft, 2019-03.

Putzer, H.J. (2020a). Ein Referenzmodell für vertrauenswürdige KI: Vorstellung eines neuen VDE-Standards - Vorstellung“, VDE tec summit, Berlin 2020-02.

Putzer, H.J. (2020b). Ein Referenzmodell für vertrauenswürdige KI: Vorstellung eines neuen VDE-Standards—Embedded Systems“, VDE tec summit, Berlin, 2020-02.

Putzer, H.J. , Nigam, V., Brunner, Th. (2019). “Dependable Autonomous/Cognitive Systems”, VDA Automotive Sys, Potsdam, 2019-06.

Putzer, H.J. , Rueß, H., Koch, J. (2021). Trustworthy AI-based Systems With VDE-AR-E 2842-61 , (in press, ID10334) embedded world 2021

VDE-AR-E 2842-61 - Design and Trustworthiness of autonomous/cognitive systems, 2020.

Wozniak, E., Cârlan, C., Acar-Celik, E. and Putzer, H.J. (2020): A Safety Case Pattern for Systems with Machine Learning Components. In International Conference on Computer Safety, Reliability, and Security (pp. 370-382). Springer, Cham.

Wozniak, E., Putzer and Cârlan, C. (2021). AI-Blueprint for Deep Neural Networks. In SafeAI@ AAAI.

2.2 Knowledge-Augmented Machine Learning

Authors:
Dr. Julian Wörmann, Dr. habil. Hao Shen , Alexander Sagel, Amit Sahu

In the era of big data, manual data analysis and decision making has become increasingly difficult due to the extreme amounts of data to be processed, the accompanying high dynamics in data streams and the ever increasing expectations for rapid provision of results. Relevant examples are omnipresent in our daily lives; predicting machine downtime based on sensor signals, recommending movies based on ratings and ranking search engine query results to name just a few.

In this context, data driven machine learning models—first and foremost deep learning architectures like artificial neural networks - have led to impressive results in various fields even exceeding human performance in accuracy and speed[6]. This performance can be attributed to the ability of these models to extract correlations in the training data that might be hidden to the human observer, either due to complex relationships between different features or just because of the sheer quantity of samples.

Although showing this remarkable performance, many of these neural networks are entirely black boxes which do not allow for human inspection of the decision process. As a consequence, this lowers the trust in these models, and even more serious, may lead to unpredictable behavior of the model in safety-critical situations.

On the other hand, deep learning models are data hungry in the sense that they can only capture correlations if a sufficient number of examples is available. For many artificially-created problems, access to data samples might not be a big deal. However, if we think of real world problems, things are changing tremendously. As an illustrative example, let's consider autonomously driving cars that should be able to perceive their environment including road demarcation, traffic and pedestrians. It is intuitively clear that it is hardly possible to cover all conceivable traffic scenarios in the training set. Moreover, uncommon situations like a ball suddenly rolling onto the street followed by a child are extremely underrepresented—not least for ethical reasons that prohibit acquiring samples that represent critical situations that may harm the environment or even human beings.

In order to address the aforementioned drawbacks of a fully data driven approach, researchers in the field of machine learning have focused on developing models that incorporate additional information ranging from common knowledge in the field of application, up to human's expertise in certain domains[7]. This way, models are expected to better adapt to specific tasks, which in turn will result in a higher robustness with regard to any unintended behavior of the model. For example, the laws of physics must not be overruled for making specific

[6] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016.

[7] L. von Rueden, S. Mayer, J. Garcke, C. Bauckhage, and J. Schuecker, Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning, Learning, 18:19–20, 2019.

outputs of a model irrelevant since they cannot be realized in practice[8] (a moving car cannot immediately stop).

Combining the strength of separate models into one approach is another strategy in order to burst the black box attitudes enabling better control of the models' behavior. Not to mention the progress in the field of explainable AI[9] with its concepts towards decomposing a neural net into its relevant components in order to validate and demonstrate trustworthy behavior in the output that is consistent with application-related knowledge.

Research challenges and results

In general, knowledge about the data, the task and the application domain is indispensable in order to build reliable and powerful Machine Learning models that can be used in practice. Starting from the identification and categorization of recent and established developments concerning the intersection of knowledge and machine learning, in the research line of Knowledge Augmented Machine Learning, we were able to achieve various results.

- Based on the insights we gained, we formulate new research challenges that pave the way to machine learning models that exhibit more human like behavior and that gather knowledge on their own.
- Regarding the augmentation of models, we were able to make visual process modelling more robust and accurate, a property that is very useful in order to predict the next states of a video sequence.
- Last but not least, safety evidence for neural networks could be demonstrated successfully, which is another building block on the way to reliable artificial intelligence.

From informed to knowledgeable models

As already mentioned in the introduction, identifying task relevant correlations in data is one of the prerequisites that allow artificially neural networks to achieve human level performance. However, this circumstance makes many models highly focused on a very specific problem they are able to solve. For example, a model that is trained to classify objects in images into different classes is unable to make any proposition with regard to semantically related tasks such as detection of bounding boxes or pose estimation. Even more, only minor changes in the data, such as variations in lighting conditions, changes in appearance or shape of the objects, might lead to undesirable performance losses of the model. As a consequence, changes in the distribution of the data or slight modification of the task usually requires a retraining of the whole model to account for these variations and to provide reliable results.

Augmenting purely data driven models with existing domain-specific knowledge is certainly key in order to establish trustworthy and reliable models that generalize well to varying conditions—intended or unforeseen—in the input data stream. To even go one step further, we follow the idea of creating models that by themselves guarantee a certain degree of versatility with respect to the domain or problem at hand. In other words, these approaches are able to acquire

[8] M. Raissi, P. Perdikaris, and G. Karniadakis, Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations, arXiv:1711.10561, 2017.

[9] S. Bach, A. Binder, G. Montavon, F. Klauschen, KR Müller, et al., On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLOS ONE 10(7), 2015.

a certain degree of knowledge about the data or task itself such that this information can be transferred or generalized to related problems as well.

We have identified invariance (Sagel et al, 2020a) as one of the key future attributes that allow models to increase the generalization capabilities. In particular, invariance to the following types have the potential to define new milestones in the era of deep learning and knowledge-based AI.

- Invariance to the skill: Deep learning can be considered function approximation between input samples and output values. Especially regarding supervised learning tasks, such as is the case in classification, where the learned function assigns labels to the input data, the flexibility in choosing the most appropriate function to achieve this task is one of the major strengths of deep learning models. However, this comes at the cost that any kind of correlation between the data and the labels can be utilized by the models, neglecting more descriptive attributes like color, texture and shape that could be useful for some other tasks such as segmentation as well. Invariance to the tasks or skill makes the trained model more versatile, which in turn would avoid unnecessary retraining and deployment of different models.
- Invariance to the data distribution: Usually it is assumed that training and test data originate from the same distribution. However, conditions might change over time, resulting in test data that differ with regard to situation, context or environment compared to the training data[10]. Models that can cope with these changes demonstrate a capability which is often referred to as out of distribution (OOD) generalization. While it is difficult to evaluate the OOD capabilities of the learned model, achieving this invariance offers promising perspectives in keeping a constant high functional quality in different situations.
- Invariance to the data syntax: Many impressive results of deep learning models have been achieved on data modalities that allow for a convenient representation such that they can be easily processed by the model or algorithm. First and foremost the processing of visual and sequential data like images or text have been the backbone of many success stories. However, many real world problems deal with structured or compositional data types like tables, graphs or sets which are much harder to process with current architectures. Transferring properties that have proven useful for vectorial data to these data types is another challenge that will significantly expand the application domains of data-driven models.

Augmenting the architecture of neural machines

Visual processes such as motion captures of crowd movements are considered random processes from a statistical point of view. With an appropriate model at hand, we can describe these visual processes in terms of their probabilistic properties. Usually, a simple linear dynamic system (LDS) model is used in order to describe the state transitions. However, many real world visual phenomena are entirely non-linear. We were able to tackle this problem by augmenting the architecture of a neural net—more precisely a variational autoencoder (VAE) —with a linear layer that models the temporal transitions (Sagel et al, 2020b). In other

[10] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, Invariant risk minimization, arXiv preprint arXiv:1907.02893, 2019.

words, our joint learning framework allows to simultaneously learn a non-linear observation as well as a linear state model from sequences of frames. In this way, the proposed model exceeds the performance of recent state-of-the-art methods in the field such as spatial-temporal generative convnet[11] and the dynamic generator model[12] evaluated on certain benchmarks. Furthermore, our model allows for a frame-by-frame synthesis of videos, which is beneficial in terms of memory and time consumption constraints.

Potential applications range from trajectory estimation[13] to anomaly detection[14], which are both essential ingredients in, for instance, motion prediction of pedestrians. Eventually, by combining two established concepts, we integrated knowledge about certain scene properties into a new model.

Safety evidence via knowledge extraction

Apart from the strategy to integrate existing knowledge into neural networks, such as via constraints, penalties or architecture modifications, safety-critical applications in particular also require strategies to validate the intended behavior of a model. In this context, formalized knowledge in terms of rules plays an important role to enable effective creation of safety evidence. Demonstrating compliance to existing knowledge increases confidence and trust in methods from the field of artificial intelligence. In order to achieve these things, we consider safety evidence as information or artefacts that contribute to developing confidence in the safe operation of the AI system[15]. In the context of this research line, we investigated the suitability of rule extraction methods[16][17] to create safety evidence for neural networks (Beyene et al, 2020). To be more precise, rule extraction methods are applied to trained neural networks along with the used training material with the goal of creating comprehensible statements about the behavior of the model in terms of simple rules. Secondly, we identified robustness to adversarial noise as a guiding safety property as it is a very crucial issue in many real world problems. In our research, we could show that surrogate models like random forests based rule extraction on the one hand provide high fidelity to the original neural network model, while on the other hand have high guaranteed safety against adversarial perturbations in the input data. Based on these findings, we can show for the first time that rules can be used as safety evidence artefacts, which allows assessing important properties of a neural network such as robustness and reliability.

Application in publicly funded industry projects—KI Wissen

In our role as a transfer institute, investigating the feasibility of our research findings in real world tasks is an objective that we pursue right from the beginning. With the start of the three-year BMWi funded project KI Wissen—which is one of four closely interconnected projects of the VDA lead initiative "autonomous and connected driving" from the artificial intelligence and machine learning in the automotive environment family of projects—we are able to validate, review and adapt our solution strategies based on use cases that occur in our daily lives.

The overall goal of KI Wissen is to combine modern data-based machine learning approaches with different types of knowledge. Even more, the extraction

[11] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, Invariant risk minimization, arXiv preprint arXiv:1907.02893, 2019.

[12] J. Xie, R. Gao, Z. Zheng, S.-C. Zhu, and Y. N. Wu, Learning dynamic generator model by alternating back-propagation through time, arXiv preprint arXiv:1812.10587, 2018.

[13] M. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R Datta, Composing graphical models with neural networks for structured representations and fast inference, in Advances in neural information processing systems, 2016, pp. 2946–2954.

[14] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, Anomaly detection in crowded scenes, in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010, pp. 1975–1981.

[15] J. L. de la Vara, et al., Towards a Model-Based Evolutionary Chain of Evidence for Compliance with Safety Standards, Sassur Workshop, SAFECOMP 2012.

[16] K. K. Sethi, D. K. Mishra, and B. Mishra, Extended Taxonomy of Rule Extraction Techniques and Assessment of KDRuleEx, International Journal of Computer Applications, 50(21):25–31, July 2012.

[17] R. Setiono, and W. K. Leow, FERNN: An algorithm for Fast Extraction of Rules from Neural Networks, Applied Intelligence, vol. 12. 2000.

of new concepts, like recurring action patterns executed by neural networks, will further increase the understanding of these models, diminishing common black box behavior. We intend to examine the potential of knowledge augmentation by means of recent developments in the field of representation learning, such as self-supervised learning[18], meta learning[19], or attention mechanisms[20]. As a result, significant impacts on issues like (1) robustness of the augmented model to unexpected or noisy inputs, (2) time/cost reduction due to smaller data sets that have to be processed, (3) interpretability of internal information processing in neural networks, and (4) reliability especially in unpredicted and safety-critical situations, are expected to pave the way towards safe and reliable autonomous driving.

Impact on industry/society and future perspective

Insights and results from this research line will have a direct impact on projects across various application domains. The aforementioned key issues concern trust and usability of AI-supported functionalities in nearly all aspects of our daily lives, ranging from healthcare, commerce, mobility and insurance, all the way to society. Endowing machine learning models with additional capabilities such that they do not only represent correlations, but also show skills that we would describe as knowledgeable, would offer entirely new application possibilities. Especially when data is scarce, when dealing with different social backgrounds or individual preferences or whenever multi-tasking qualities are desired, the designing of knowledge augmented machine learning models will offer vast potential that needs to be leveraged in future research.

[18] L. Jing and Y. Tian, Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey, In IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI), 2020.

[19] C. Finn, P. Abbeel, and S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, In Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1126–1135. JMLR. org, 2017.

[20] S. Zhang, J. Yang, and B. Schiele, Occluded pedestrian detection through guided attention in CNNs, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6995–7003, 2018.

LITERATURE

Beyene T.A., Sahu A. (2020): Rule-Based Safety Evidence for Neural Networks. In: SAFECOMP 2020 Workshops.

Sagel, A., Sahu, A., Matthes, S., Pfeifer, H., Qiu, T., Rueß, H., Shen, H. and Wörmann, J. (2020a). Knowledge as Invariance—History and Perspectives of Knowledge-augmented Machine Learning. arXiv preprint arXiv:2012.11406.

Sagel A., and Shen, H. (2020b). Dynamic Variational Autoencoders for Visual Process Modeling. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3677–3681, Barcelona, Spain, 2020.

2.3 Joint Action Planning

Authors:
Klemens Esterle, Patrick Hart, Tobias Kessler

With robotic systems (drones, cars, or medical robots) being developed with increasing capabilities to act autonomously, they will usually rely on a recognize-act-cycle (sense-plan-act or perception-deliberation-execution) as shown in Figure 6. Specifically, the deliberation module calculates the system's action based on its prior or perceived knowledge. As part of the deliberation component, joint action planning is an advanced research line dealing with action planning in multi-agent settings, such as systems with multiple interacting intelligent agents (humans or autonomous systems). These methods find the robot's best action (which we will denote as "ego agent") by planning, predicting, and evaluating all other agents' actions and reactions in conjunction with the ego agent. Other agents may interact with the ego agent, for example, in a cooperative way (swarms of service drones) or a more competitive manner. If the action plan does not account for these interactions accurately, the overall system's safety is at risk.

When engineering autonomous systems, the system requirements that need to be defined are manifold. Generally speaking, autonomous systems are expected to operate safer than humans. Broken down to the deliberation component,

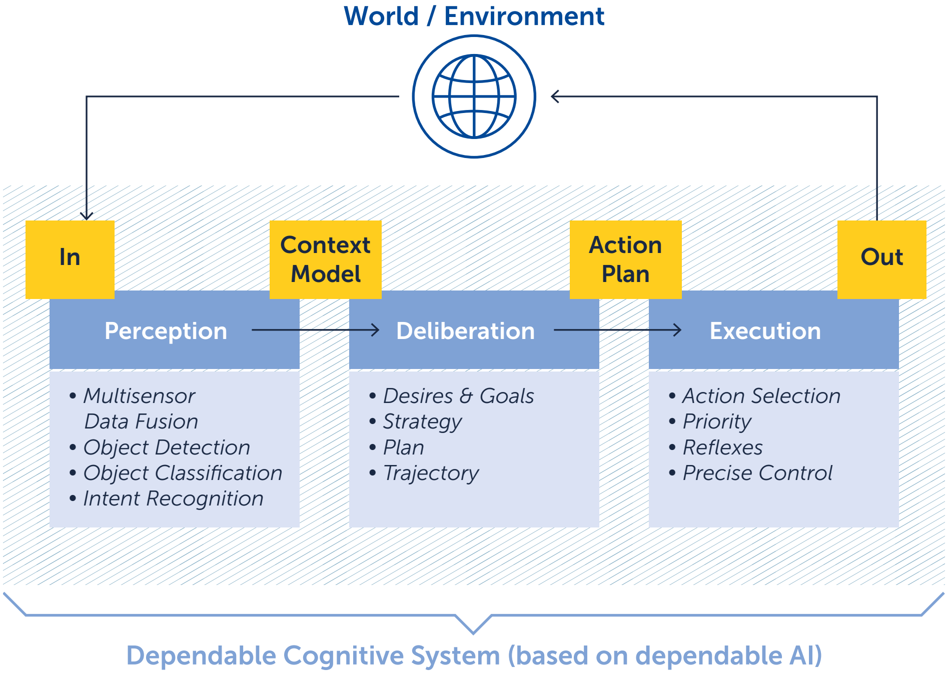


Figure 6.
Full cycle of autonomous systems that captures the perception, the deliberation and the execution.

this means that a joint action plan must be free of any collision despite various model uncertainties and faults. Apart from gaining the acceptance of customers and society, an autonomous system must balance safety and efficiency when trying to achieve its mission goals. For example, a self-driving vehicle needs to blend into traffic and not stand on a lane for minutes waiting to merge (the so-called "freezing robot problem"). Often, another requirement calls for autonomous systems to be rational and interpretable by humans.

Part of this may require the joint action planner to consider the rules and norms of the environment. Of course, this list could go on, but it essentially shows the tight coupling of requirements between the system and deliberation level. The techniques that will realize those capabilities must be designed to satisfy the product's full operational design domain. For example, autonomous vehicles need to generalize to various cities and countries, environmental conditions, and the behavior of other human drivers, cyclists and pedestrians.

AI-based embedded systems such as swarms of service drones, autonomous vehicles, or surgical robots will need to apply joint action planning techniques, depending on their operational design domain and integration level, to operate side-by-side with humans.

Current variants of autonomous vehicles have not reached market readiness for a variety of reasons. One of the core difficulties is the targeted operation of autonomous cars side-by-side with humans. For these, models are being built and used to approximate their behavior. Making efficient use of these human agent models lies within the domain of joint action planning, as we discussed before, where a planner models the uncertain interaction with other traffic participants by planning a joint action for the ego vehicle and the surrounding vehicles. Game-theory offers fundamentally sound and mature concepts to achieve this. Besides safety, the acceptance of these systems will depend on customer satisfaction (comfort, mission completion, etc.). For this, autonomous cars need to find the best possible action by considering the reactions based on several potential actions. Further, these must also be capable of learning from experience how to blend into mixed-traffic over time—be it with other autonomous vehicles or human traffic participants.

Research challenges and results

In terms of deliberating AI-based self-driving cars, significant progress has been made in recent years. However, various challenges remain to be solved, such as the robustness against uncertainties and faults, the methodological examination of these, and validating the safety of AI-based deliberation components.

Handling Uncertainties

As stated above, uncertainties are omnipresent for many reasons, such as sensor limitations, distribution shifts from simulation to the real world, and many others. These uncertainties concern models built from observed data at runtime ("online" models) and models constructed from knowledge or data in the development phase ("offline" models). With online models, such as perceived environment models, imprecision might originate from localization or object detection. Quantifying uncertainty is an active field of perception research that heavily relies on

AI-components. Novel planning approaches using the obtained uncertainties in joint action planning are required to generate safe behaviors.

In contrast, offline models incorporate knowledge from the development engineer, e.g., vehicle models, environment models (Esterle et al, 2020b), and traffic participant models that guide and restrict the solution algorithms at runtime. Early results of our work at fortiss are promising in terms of robustness against online model inaccuracies (Kessler et al, 2020). A combination of coverage criteria for offline models with online model refinement increases the robustness against distribution shifts (Bernhard et al, 2020). Even further, for autonomous systems to work well in various scenarios and countries and to adjust to new situations, new and observed data might be used to update or correct the offline and online models. At fortiss, we also developed a framework for evaluating learned behaviors that use deep neural networks as approximation functions against model inaccuracies at runtime (Hart et al, 2020). This enables the system to evaluate potentially catastrophic distribution shifts at runtime and helps to bring insights into the generalization capabilities of the learned behavior policy. Eventually, when self-driving vehicles shift from prototype to product, they will need to be robust against imprecision in online and offline models.

Methodological Examination

To support the robustness argumentation in a safety assurance case for obtaining certification, the AI-based components need to be evaluated in terms of their robustness, how well these generalize over various scenarios, how these handle known unknowns and unknown unknowns, and more. At fortiss we develop an examination and verification framework for AI-based components called BARK, which enables the modeling of the aforementioned uncertainties and distribution shifts individually for each vehicle. The tool is fully open-source (refer to BARK). It serves as an ideal platform for developing and evaluating novel AI-based deliberation components. Details on the first version and the benchmark and verification capabilities can be found in Bernhard et al, 2020.

Verifiably safe deliberation

With the configuration space being vast, full coverage of the scenario space that considers all types of conditions, uncertainties, and faults cannot be guaranteed in an offline evaluation. We argue that to verify an AI-based deliberation component at runtime, planning and monitoring methodologies must be developed side-by-side. Recently, there has been a strong effort towards verifiable deliberation components by introducing the responsibility-sensitivity safety metric (RSS[21]) or by employing reachability theory in real-time[22]. Although these methods provide provably safe behaviors, they fall short in blending seamlessly into mixed-traffic as they assume worst-case behavior of other traffic participants. Thus, AI-based deliberation components have to be designed to act preventatively to the safety monitor so that the safety monitoring concept never or rarely needs to engage. For this, inherently safe-by-design solution methods are required that can provide safety guarantees at runtime, such as safe-reinforcement learning or well-established optimization techniques as we have shown in Esterle et al, 2020a. Based on optimization techniques that have convergence guarantees, we also evaluate what the price for this optimality is in terms of algorithmic perfor-

[21] Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a Formal Model of Safe and Scalable Self-driving Cars. ArXiv, 1–37. https://arxiv.org/pdf/1708.06374.pdf

[22] Pek, C., Manzinger, S., Koschi, M., & Althoff, M. (2020). Using online verification to prevent autonomous vehicles from causing accidents. Nature Machine Intelligence, 2(9), 518–528. https://doi.org/10.1038/s42256-020-0225-y

mance and computational efficiency. We also present results that show how driving styles can be safely adopted on the observed behavior of other agents (Kessler et al, 2020). However, additional approaches are required that are capable of learning behavior policies and that can blend into mixed-traffic, such as combining learned behaviors with an inherently safe optimization (Hart et al, 2019).

Benefits

The Joint Action Plannin research line at fortiss develops safe-by-design and certifiable deliberation components for solving even the most critical and difficult driving situations. Using concepts from joint action planning, novel deliberation components are evaluated for their use within safety-critical applications in a methodological manner and to demonstrate the overall certification concept. Using our tailored simulation tool BARK and our autonomous driving research vehicle fortuna (cf. Section 3.1.4), we show the applicability in real-road driving scenarios. The developed methodologies are not restricted to autonomous driving but shall be transferred to various domains, such as drones and logistic robots. By doing so, a holistic view of AI-based deliberation systems is developed that takes the full life cycle of AI-based deliberation components into account from development to deployment.

LITERATURE

Bernhard, J. & Knoll, A. (2020). Robust Stochastic Bayesian Games for Behavior Space Coverage. arXiv preprint arXiv:2003.11281.

BARK: https://github.com/bark-simulator/bark.

Bernhard, J., Esterle, K., Hart, P., & Kessler, T. (2020). BARK: Open behavior benchmarking in multi-agent environments. arXiv preprint arXiv:2003.02604.

Esterle, K., Kessler, T., & Knoll, A. (2020a). Optimal Behavior Planning for Autonomous Driving: A Generic Mixed-Integer Formulation. In 2020 IEEE Intelligent Vehicles Symposium (IV) (pp. 1914-1921). IEEE.

Esterle, K., Gressenbuch, L., & Knoll, A. (2020b). Modeling and Testing Multi-Agent Traffic Rules within Interactive Behavior Planning. IROS 2020 Workshop on Perception, Learning, and Control for Autonomous Agile Vehicles.

Hart, P., & Knoll, A. (2020). Using Counterfactual Reasoning and Reinforcement Learning for Decision-Making in Autonomous Driving. IROS 2020 Workshop on Perception, Learning, and Control for Autonomous Agile Vehicles, 2020.

Hart, P., Rychly, L., & Knoll, A. (2019, October). Lane-Merging Using Policy-based Reinforcement Learning and Post-Optimization. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (pp. 3176-3181). IEEE.

Kessler, T., Esterle, K., & Knoll, A. (2020). Linear Differential Games for Cooperative Behavior Planning of Autonomous Vehicles Using Mixed-Integer Programming. In 2020 59th IEEE Conference on Decision and Control (CDC) (pp. 4060-4066). IEEE.

2.4 Verification of Machine Intelligence

Authors:
Dr. Mojdeh Golagha, Amit Sahu

Considerable effort has been made in recent years to enable machine intelligence by applying AI in different systems; to the point that AI systems are becoming pervasive in our life. Health care, financial systems, insurance, autonomous transportation, and many other areas have been influenced by AI. Despite their success, a fundamental challenge remains: to ensure that AI-enabled systems behave as intended. This challenge has become critical in cases where the performance of such systems is critical, such as autonomous driving, where an accident can happen and lead to fatalities. Therefore, there is an increasing need for new methodologies to test and verify these kinds of systems. By testing, we evaluate the AI-enabled system in different conditions, observe its behavior, and look for faults. With verification, we produce the argument that the system will not misbehave under these different conditions.

We test and verify AI-enabled systems at two levels: component level and system level. Component testing is the testing of specific components of a product. It is usually done in isolation from the rest of the components. In system level testing, the components are tested as a whole to ensure that the overall product meets the specified requirements. At this level, the product is tested in an environment that is very close to that which the user will experience once it is deployed. For instance, to test an autonomous vehicle (AV), among other steps of testing, we need to test components (DNN component responsible for pedestrian detection), and subsystems or advanced driving assistant systems (ADAS) (emergency braking system), and finally the entire AV.

Component level

At the component level, we focus on one specific component that is handled by ML, such as a vision-based perception component handled by a neural network YOLO[23]. These learning enabled components are difficult to verify with traditional methods. Uncertainties inherent in the ML algorithms, because of their data-driven approach, limit their integration into a system, especially in safety critical domains. Specifically, for ANN-enabled self-driving vehicles it is important to establish properties related to the resilience of ANNs to noisy or even maliciously manipulated sensory input. Additionally, in the absence of best safety engineering practices for NN, there is an urgent need for an adequate set of metrics for measuring all important dependability attributes. Safety for autonomous vehicles needs to be systematically tested for models learned from neural networks.

[23] Alexey Bochkovskiy, Chien-Yao Wang, and H. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” in ArXiv, vol.abs/2004.10934, 2020.

Research challenges and intermediate achievements

There have been constant efforts to make the ML components dependable. Ethics guidelines for trustworthy AI have been published by the European Commission[24]. Software frameworks—like TensorFlow, PyTorch—have integrated and introduced verification and explanation tools. However, despite huge efforts and partial solutions, the field is still suffering from unsolved challenges.

Noise resilience

One of the biggest threats to verify and certify neural networks is their vulnerability to adversarial noise. Specifically, by addition of a non perceivable (to humans) deliberate noise, neural networks can be fooled to change their classification decision. The solutions offered by the NNs dominate most of the traditional methods in terms of efficiency and effectiveness. However, this vulnerability, even if it can only occur by manual effort, reduces the trust in systems built with these components.

In contrast to the direct approach of making the NNs robust to adversarial noise, we targeted the problem by defining resilience properties of ANN-based classifiers. The aim was to provide a measure of the robustness against the noise. This was done by using formal logic and SMT solvers to find the maximum amount of input or sensor perturbation which is safe and within the decision boundary of the neural network model (Cheng et al, 2017a).

Dependability attributes

Neural networks are increasingly being applied to more and more applications. With the SOA results in perception (YOLO), they are becoming quintessential components in safety-relevant applications such as highly-automated driving. However, state-of-the-practice safety engineering processes (cmp. ISO 26262) require that safety-relevant components, including NN-enabled ones, must satisfy their respective safety goals. Directly applying traditional testing methods and corresponding test coverage metrics such as MC/DC (cmp. DO 178C) to NNs may lead to an exponential (in the number of neurons) number of branches to be investigated.

Neural networks are vastly different in their workings from software codes—a single neuron activation is not strongly connected to the result of the network. Hence, essential dependability aspects like robustness, interpretability, correctness, and completeness—RICC—needs to be redefined for the NN components.

To keep such methods tractable on deep neural networks (SOA architectures), we redefined and developed metrics that attempted to approximate the dependability aspects. These metrics were NN-specific and efficiently computable (Cheng et al, 2018a).

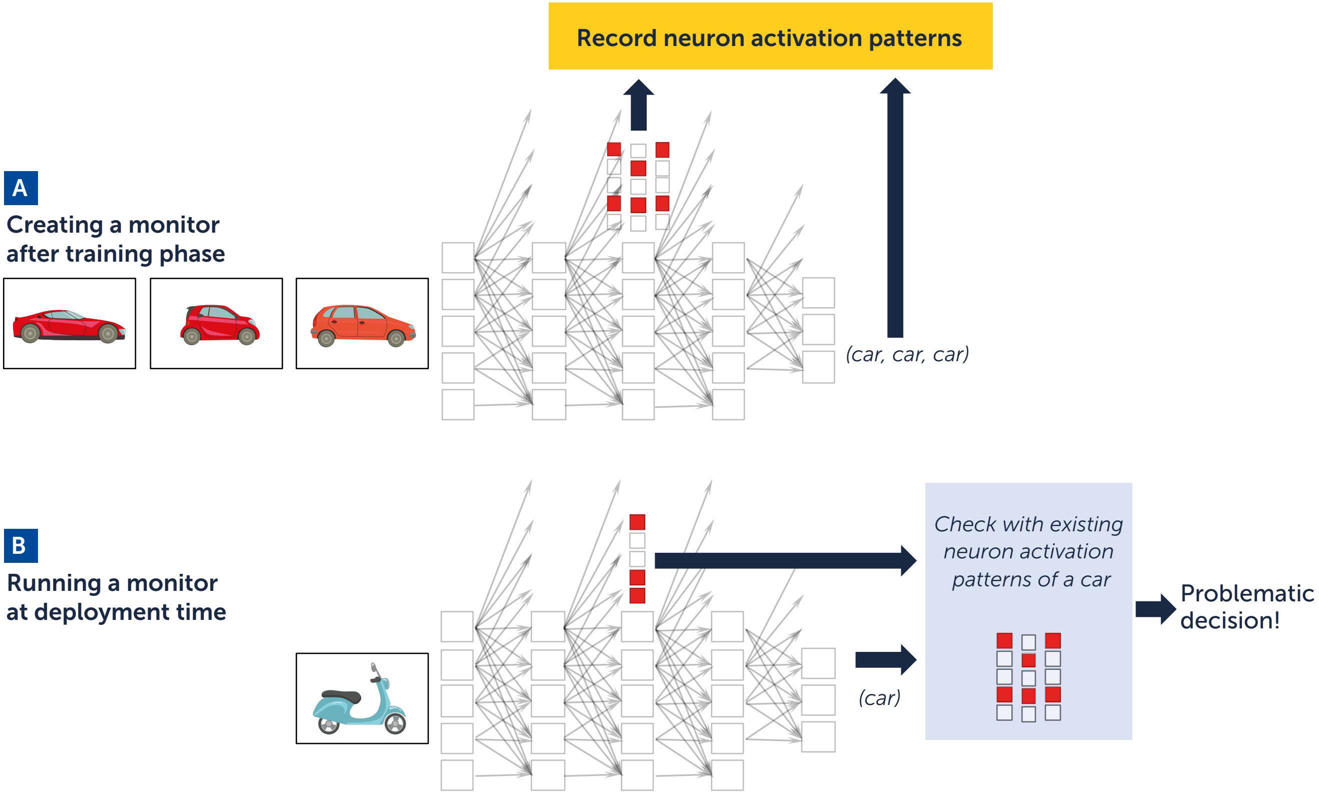
[24] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Runtime monitoring

Runtime monitoring is a broad concept. One could monitor specific behavior, known errors or build a safety envelope over the ML component. There have been vast contributions towards each of these directions. In contrast, our aim was to identify if a decision made by a neural network is supported by prior similarities in training (Cheng et al, 2019).

Neural networks are data-driven algorithms. Therefore, their application is bounded to the dataset distribution they were trained on. However, NNs always output a confidence score, no matter the dataset. This can be misleading, for example, when searching for a car type and getting a score for a motorbike. Hence, runtime monitoring was defined in a way to measure the distance of the given data point from the distribution of the training dataset. In other words, a motorbike data point will be compared to the distribution of the cars dataset as shown in Figure 7.

Figure 7.
Runtime monitoring using neuron activation pattern.


Record neuron activation patterns
A
Creating a monitor after training phase
(car, car, car)
B
Running a monitor at deployment time
Check with existing neuron activation patterns of a car
(car)
Problematic decision!

Software

All these approaches and more were combined into one central toolkit—Neural Network Dependability Kit (NNDK). NNDK is available as an open source tool on GitHub (Cheng et al, 2019b). The software supports developers in handling the dependability aspects of the neural networks.

NNDK application

Our approach is different from individual metrics in that we offer verification (via metrics) during each stage of AI development. A development cycle goes through the following stages: 1. Data Preparation 2. Training and Validation 3. Testing and Generalization and 4. Operation.

In the following, we describe how NNDK covers each stage:

1. **Data Preparation**
   Data collection is the most basic and essential task as every other phase is affected by its quality. High-quality datasets cover all scenarios and situations. However, obtaining a complete set of scenarios can result in a combinatorial explosion. Hence, to quantify the coverage of the scenarios by the dataset, NNDK offers scenario k-projection coverage (Cheng et al, 2018b).

2. **Training and Validation**
   Formal reasoning ensures that the model satisfies the risk properties to ensure predictable/reliable behaviour under conditions that are similar, but possibly different, to the ones experienced in the test cases.

3. **Testing and Generalization**
   NNDK offers neuron k-projection coverage over a preselected layer. This measures the completeness of the test set to cover the whole neuron layer under analysis. Also, NNDK offers a perturbation loss metric to measure how the system performs in noisy environments.

4. **Operation**
   One can only expect adequate performance from the NN model when it is applied to a data point with prior similarities to the training data. NNDK keeps track of data points by recording their Neural Activation Pattern (NAP) on a preselected layer.

Results and the process of applying the tool set is available as a research paper on ArXiv. Practical applications of NNDK into the development process of two use cases—Diabetic Retinopathy (DR) classification, Smart Tunnels— were done as part of the FED4SAE project.

Case study 1: diabetic retinopathy

The objective of the development was to create a working prototype that demonstrates the classification of retinal fundus images for the presence of diabetic retinopathy (DR) indicators.

Overall results of integrating NNDK in the development process:

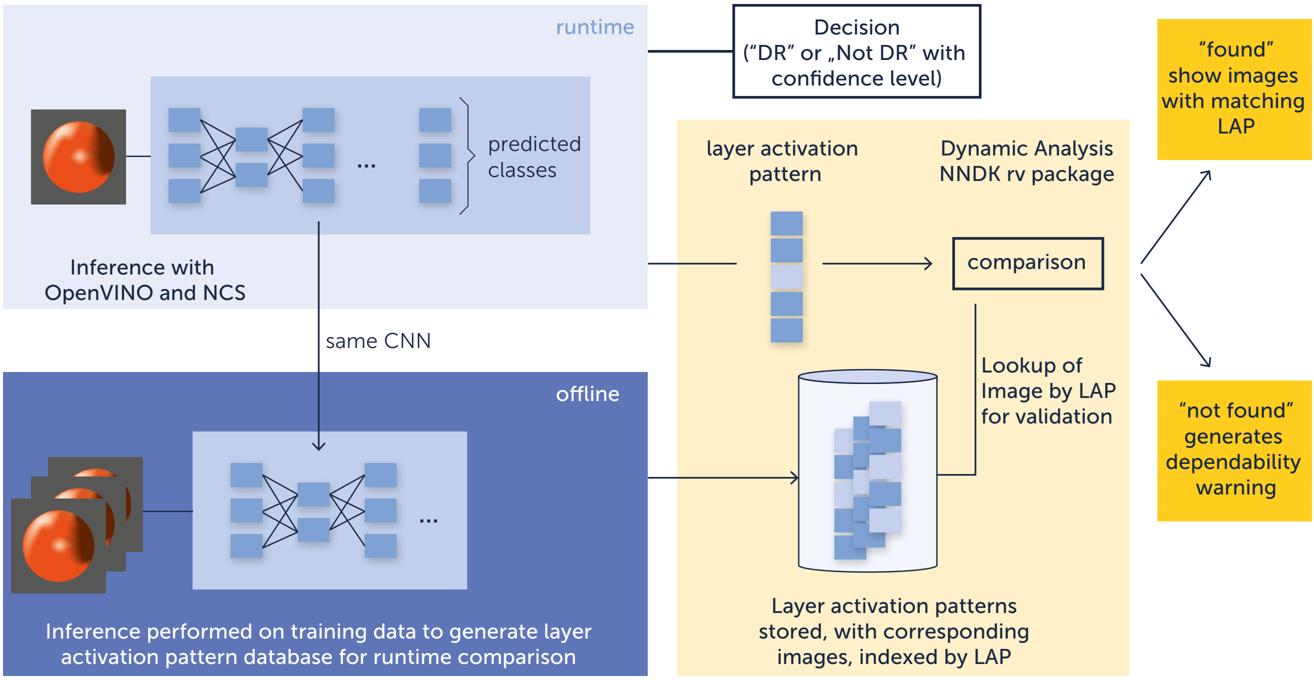
Static analysis

- Increase in accuracy of the final model due to refined application and structure of the model architecture 69% – 77%
- More understanding about the model's features
- Developers were able to identify activation of arbitrary features from ImageNet in the old model vs disease relevant features in the newly trained model.
- Pruning was suggested by the dependability metrics.
  Result: 74% pruning of neurons had only 0.1% accuracy drop

Dynamic analysis

NN decision was supported by prior similarities in the training data. Retrieval of corresponding images from the training dataset for further analysis (build trust) from doctors. Figure 8 shows the architecture for applying the runtime monitoring in the DR application. In the final output, either cases similar to the current patient were retrieved or a dependability (out of distribution) warning was issued.

Figure 8.
Runtime monitoring on DR application


runtime
predicted classes
Inference with OpenVINO and NCS
Decision ("DR" or „Not DR" with confidence level)
"found" show images with matching LAP
layer activation pattern
Dynamic Analysis NNDK rv package
comparison
same CNN
offline
Lookup of Image by LAP for validation
"not found" generates dependability warning
Inference performed on training data to generate layer activation pattern database for runtime comparison
Layer activation patterns stored, with corresponding images, indexed by LAP

Case study 2: smart tunnels

The objective of the development was an automatic incident detection (AID) system for road tunnels using neural networks. NNDK gave insights that directed the development and selection of the NN model:

- Neuron k-projection coverage and NAP metrics showed that very few neurons were activated for both classes (pedestrian, stationary vehicle). This suggested that network pruning would be a good next step. This was an essential insight as efficiency of the inference step was a key requirement in this use case due to the high number of images that need to be processed continuously.

- Applying the perturbation loss metric, it was discovered that the images were highly prone to the noise, resulting in a 95% average loss of confidence. To deal with this problem, dataset augmentation was done and new models were trained. The model with the highest score was finally selected.

Benefits

NNDK offers metrics to measure dependability attributes—robustness, interpretability, correctness, completeness—in the neural network. The integration of these metrics in the development process has been validated on research datasets and two practical use cases. The case studies demonstrate the usage of the dependability kit to obtain insights into the NN model and how they informed the development process of the neural network model. After interpreting neural networks via the different metrics available in the NNDK, the developers were able to increase the NNs' accuracy, trust the developed networks, and make them more robust.

System level

A popular suggestion to test ADAS/AVs at the system level is to randomly pick test cases for virtual testing from huge mileages of pre-recorded drives. Word of mouth suggests that roughly 6.6 billion kilometers are sufficient[25].The direct reuse of recorded drives for testing purposes as well as the random approach are questionable undertakings. In terms of random testing, the search space simply is too large. In terms of directly reusing recorded drives, the quality of (recorded) test cases is system-specific. Recorded test cases may be "good" test cases for one system (version/variant) and useless for another. The general idea then is that test cases that may be "good" in that they trigger behaviors for one system, may be questionable for others in that they do not even provoke the functionality to be tested (Hauer et al, 2020a). This consideration unfortunately also implies that we cannot expect a single sensible "reference test suite" for AVs as known from other domains. Instead, system-specific test cases have to be generated.

A better approach is testing AVs in simulation using scenario-based testing where such driving systems are tested in recurring and challenging traffic scenarios. The recurring traffic scenarios are called "scenario types"[26]. One example is

[25] Wachenfeld, Walther, and Hermann Winner. "The release of autonomous vehicles." Autonomous driving. Springer, Berlin, Heidelberg, 2016. 425–449.

[26] T. Menzel, G. Bagschik, and M. Maurer. Scenarios for development, test and validation of automated vehicles, arXiv, 2018.

a vehicle following another on the right lane of a two-lane highway when both vehicles are overtaken by a third vehicle. During testing, scenario types are used to generate concrete scenario instances (Hauer et al, 2020b). These scenario instances are in fact test cases. In the example, different instances may consider different driving speeds or distances between cars. The goal of scenario-based testing is to identify instances that stress the autonomous driving behavior (near-crashes, abrupt acceleration, or deceleration) (Hauer et al, 2020b). Proving that the system works as expected in the challenging instances increases confidence in the system (Hauer et al, 2020b). To be able to generate test cases, the first requirement is having a "complete" list of scenario types. However, achieving "completeness" is challenging.

A common approach in industry is to have experts manually create such lists of scenario types. However, the manual creation of such catalogs poses risks on the completeness and adequacy of the list (Hauer et al, 2020b). Since experts use their mental models to define scenario types, some scenario types might be overlooked. Also, the granularity they consider to define a scenario type might not be correct (Hauer et al, 2020b).

To improve the quality of scenario types lists and augment the manual scenario creation by experts, we proposed an approach to automatically extract scenario types from real recorded driving data. We did our first experiment on highway driving data[27] and published the results in Hauer et al, 2020b. In this work, we extracted scenario types from real driving data by clustering recorded scenario instances, which are composed of time series. The distance between the ego vehicle and all its surrounding vehicles form the dataset used for clustering. Next, we extended our clustering idea to extract scenario types for roundabouts[28] and Intersections[29].

We have inferred more than 100 clusters/scenario types. On a more foundational level of research, organizing the inferred scenario types, we generated a living hierarchical set of scenario types.

The *hierarchical* system represents scenario types at three levels of granularity (road type, trajectory of ego vehicle, and distance to surrounding vehicles). The *living* system of scenario types can change over time as we gather more data. Under the assumption that new recorded data is continuously available, either by test vehicles or by (near) accidents that were explicitly reported, we (1) can continuously decide if a genuinely "new" scenario type has happened in the real world (2) can add a new scenario type to existing catalogs (3) check if reclustering over the full data set needs to be done or if this can be done locally.

After collecting a high quality scenario catalog, the next step is generating scenario instances and deriving good test cases for each scenario type. To this end, we proposed a search-based technique to put the system under test under a safety-critical situation (safety distance less than the threshold) and see how it behaves. We published our results in Hauer et al, 2019.

In conclusion, we developed a novel methodology and technology for deriving tests from scenario types and technology for generating scenario types from recorded drives.

[27] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems, 21st International Conference on Intelligent Transportation Systems (ITSC), 2018.

[28] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein, rounD Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany, submitted.

[29] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein, inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections, , arXiv, 2019.

LITERATURE

Cheng, C. H., Nührenberg, G., & Rueß, H. (2017a). Maximum resilience of artificial neural networks. In International Symposium on Automated Technology for Verification and Analysis (pp. 251–268). Springer, Cham.

Cheng, C. H., Huang, C. H., & Nührenberg, G. (2019b). nn-dependability-kit: Engineering neural networks for safety-critical autonomous driving systems. In 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (pp. 1–6). IEEE.

Cheng, C. H., Huang, C. H., Rueß, H., & Yasuoka, H. (2018a). Towards dependability metrics for neural networks. In 2018 16th ACM/IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE) (pp. 1–4). IEEE.

Cheng, C. H., Nührenberg, G., & Yasuoka, H. (2019c). Runtime monitoring neuron activation patterns. In 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE) (pp. 300–303). IEEE.

Cheng, C. H., Huang, C. H., & Yasuoka, H. (2018b). Quantitative projection coverage for testing ml-enabled autonomous systems. In International Symposium on Automated Technology for Verification and Analysis (pp. 126–142). Springer, Cham.

Hauer, F., Pretschner, A., & Holzmüller, B. (2020a). Re-using concrete test scenarios generally is a bad idea. In 2020 IEEE Intelligent Vehicles Symposium (IV) (pp. 1305–1310). IEEE.

Hauer, F., Gerostathopoulos, I., Schmidt, T., & Pretschner, A. (2020b). Clustering traffic scenarios using mental models as little as possible. In 2020 IEEE Intelligent Vehicles Symposium (IV) (pp. 1007–1012). IEEE.

Hauer, F., Pretschner, A., & Holzmüller, B. (2019). Fitness functions for testing automated and autonomous driving systems. In International Conference on Computer Safety, Reliability, and Security (pp. 69–84). Springer, Cham.

2.5 Human-centered Machine Learning

Authors:
Dr. Yuanting Liu, Dr. habil. Hao Shen

Motivation

Human-centered Machine Learning (HCML) aims to design machine learning systems that empower humans, by raising their self-efficacy, promoting their creativity and respecting their responsibility, rather than replacing them. The symbiosis between humans and intelligent systems is key to demonstrably reproducing and understanding the underlying rationale of decisions made by machine learning algorithms with the intention of improving system usability and developing useful applications.

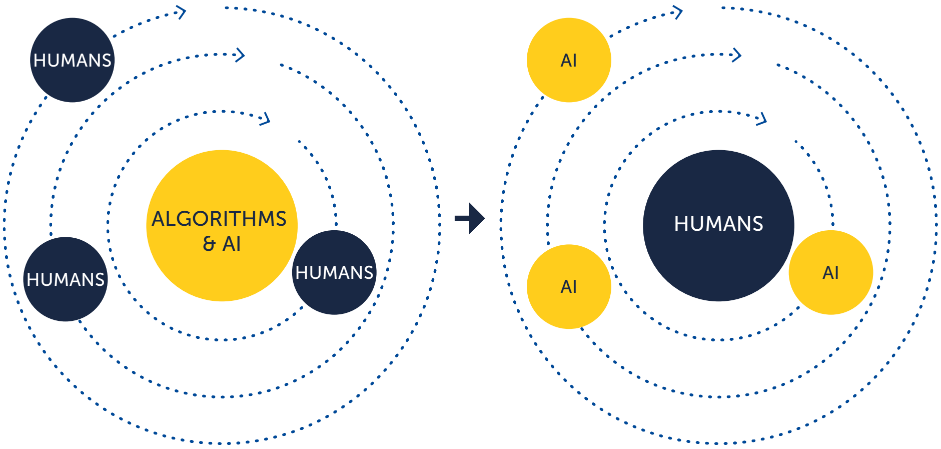


Figure 9. A second copernican revolution puts humans at the center of attention[30]

Consider, for example, the recent case of Boeing 737 Max crashes[31] and Amazon's recruiting tool that was biased against women[32] due to excessive automation and machine learning (ML) with biased datasets. The current focus on fully automatic decision support systems and explanation techniques for these systems is inherently, but unintentionally data-centric. Humans are not capable of performing well as passive supervisors, no matter how well designed the explanation interface of an ML system is. Therefore, placing humans as supervisors

[30] B. Schneidman. "Human-centered Artificial Intelligence: Three Fresh Ideas". International Journal of Human-Computer Interaction 2020, 12(3), 109–124.

[31] https://www.justice.gov/opa/pr/boeing-charged-737-max-fraud-conspiracy-and-agrees-pay-over-25-billion

[32] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

over AI systems without involving them in the actual task is an ineffective way to handle problems like intrinsic biases in its prior data. This makes it hard to overcome barriers regarding safety, ethics, and social justice to the deployment of ML systems in high-stakes applications. AI systems are required in order to make the shift from attempting to place the human into the AI and algorithm loop, to building human-centered AI-based machine-in-the-loop concepts (Figure 9). For the purpose of balancing a high level of human control and automation, the increasing need for human-centered AI and ML is becoming apparent every day[33].

Approach and results

The introduction of human factors aims to enhance ML algorithms and to improve user satisfaction while ensuring an acceptable task accuracy. To facilitate such human-machine-collaboration and -interaction systems based on trust, ML systems must be able to provide explanations when decisions or actions are made by specific ML algorithms. Our methods include user-centered design approaches (UCD), machine-in-the-loop and human-in-the-group, adapting to the user with rich human feedback, and implicit feedback, through imitation and with active learning. Human knowledge is infused into the machine-learning process with the goal of further increasing data efficiency, and improving the robustness of the learned result. By leveraging increasingly connected and autonomous systems, we develop ML systems that ensure a measurable quality of adequateness for human users while respecting human autonomy and self-determination with machine intelligence. In this way, ML techniques can be improved continuously in a safe and efficient way by reinforcing human-machine collaboration.

For this reason, one of our focuses is on research into developing user modeling and user-adaptive interaction (Schmidmaier et al, 2019; Klingner et al, 2020) and building up transparency and trust that allows users to gain insight into the system's decision (Wiegand et al, 2019a; Wiegand et al, 2019b).

By safely and efficiently controlling and improving the learning process, especially for intelligent human-in-the-loop (HitL) systems (Han et al, 2019; Weber et al, 2020), we demonstrated useful techniques for selected use cases, such as stress-detection for firefighter applications[34] at the IBM fortiss Center for AI. Firefighters are one of the most vulnerable insured working groups in the statutory accident insurance system[35]. Extreme heat, poor visibility due to smoke, time pressure, danger, all of these factors lead to immense stress, reduced situational awareness, and potentially severe impairment of cognitive abilities.

To tackle this challenge, IBM and fortiss have assembled a team to focus on the development of data-driven human-centered machine learning algorithms for stress monitoring based on data mining and cognitive characteristics. The proposed solution enables the shaping of new stress recognition models by means of various firefighting scenarios and valuable experience gained from such missions. Several candidate ML approaches were investigated to measure and estimate the stress level of firefighters in real-time with the goal of assisting mission commanders in critical decision making. Moreover, for the underlying stress data, an in-house virtual reality-based tool was also built to collect general stress indicators, such as heart rate, brain activity, muscle tension, and skin conductance as task inputs (Klingner et al, 2020). As a proof of concept, various ML approaches have demonstrated their effectiveness for developing stress detection

[33] https://www.faz.net/aktuell/wissen/klug-verdrahtet/klug-verdrahtet-endlich-licht-in-der-black-box-16198015.html

[34] More details in the Chapter 4 on "Development of improved personalized stress detection models and a VR stress simulation for firefighter to create new datasets".

[35] https://www.dguv.de/de/mediencenter/pm/pressemitteilung_402783.jsp

models, such as self-supervised learning (SSL) for "label-free" feature extractors (Matthes et al, 2021), as well as an efficient personalization method trained with physiological data and limited labels via HitL interactions.

Another use case example is a recommender system for searching office, laboratory and IT supplies, or warehouse and business equipment in collaboration with Mercateo Deutschland AG[36] for their commercial customers, as part of the IuK Bayern project HighWoWNet. To provide prediction, suggestion or rating of items in the form of a textual or image-based description to customers, a high-quality recommender system positively impacts the users' experience and the overall enterprises' revenue or decision making. Thus, it is important to choose the best recommendation algorithm such as an optimal collaborative filtering algorithm using a deep neural network method[37] [38]. Due to the shallow structure, classic graph neural networks (GNNs) failed in modelling high-order graph structures that deliver critical insights into task relevant relations. The negligence of those insights leads to insufficient distillation of collaborative signals in recommender systems. fortiss therefore proposes a unified GNN framework tailored for recommendation tasks, which is capable of automatically selecting the useful information in prior knowledge (Han et al, 2021). Moreover, this approach, which involves customers actively and passively using their preferences and behaviors in recommender systems (HitL), can help to personalize the experience, alleviate data latency and enhance scalability and performance.

Benefits

HCML augments and enhances the human experience while ensuring human supervisory influence and control of ML systems, especially in critical, high-stakes domains such as aviation, healthcare, fintech and law enforcement. Nowadays, human-centered AI and ML is becoming a highly-popular topic in industry, research and society. In this research line, the techniques that are developed always focus on the human's interest and needs. Specifically, the aim is to enable the interpretability of AI/ML solutions from the human perspective and to enhance trust betwen AI/ML systems and human users. The knowledge and methods in this research line are therefore crucial for safety critical applications, ranging from aerospace, transportation, healthcare, and many other privacy-sensitive scenarios. As AI technologies rapidly advance in both industrial and daily scenarios in the foreseeable future, it is believed that HCML will eventually form a core pillar of the fortiss AI strategy in expanding research expertise and establishing industrial influence. fortiss established this research line two years ago. We have meanwhile successfully collaborated with industrial partners and are making further contributions to national and international projects (such as EU H2020 HumanE-AI-Net, LuFo-VI KIEZ4-0, BMWi KI Wissen).

[36] http://www.mercateo.com

[37] S. Guens, K. Coussement, K.W. De Bock. "A framework for configuring collaborative filtering-based recommendations derived from purchase data". European Journal of Operational Research. 2018, 265(1), 208–218.

[38] P. Sulikowski, T. Zdzieko. "Deep-Learning—Enhanced Framework for Performance Evaluation of a Recommending Interface with Varied Recommendation Position and Intensity Based on Eye-Tracking Equipment Data Processing". Eletronics 2020, 9 (2), 266.

LITERATURE

Han, Z., Anwaar, M., Arumugaswamy, S. et al. (2021). Metapath- and Entity-aware Graph Neural Network for Recommendation, The 24th International Conference on Artificial Intelligence and Statistics. Under review

Han, Z., Weber, T. , Matthes, S. ,Liu, Y., & Shen, H. (2019). Interactive image Restoration, Human-centric Machine Learning Workshop at NeurIPS.

Klingner, S., Han, Z., Liu, Y., Fan, F., Altakrouri, B., Michel, B., ... & Chau, S. M. (2020). Firefighter Virtual Reality Simulation for Personalized Stress Detection. In German Conference on Artificial Intelligence (Künstliche Intelligenz) (pp. 343–347). Springer, Cham.

Matthes, S., Han, Z., Qiu, T. et al. (2021). Personalized Stress Detection with Self-supervised Learned Features, International Joint Conference of Artificial Intelligence 2021. Under review

Schmidmaier, M., Han, Z., Weber, T., Liu, Y., & Hußmann, H. (2019). Real-time personalization in adaptive IDEs. In Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (pp. 81–86).

Wiegand, G., Schmidmaier, M., Weber, T., Liu, Y., & Hussmann, H. (2019a). I drive-you trust: Explaining driving behavior of autonomous cars. In Extended abstracts of the 2019 chi conference on human factors in computing systems (pp. 1–6).

Wiegand, G., Mai, C., Holländer, K., & Hussmann, H. (2019b). Incarar: A design space towards 3d augmented reality applications in vehicles. In Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications (pp. 1–13).

Weber, T., Hußmann, H., Han, Z., Matthes, S., & Liu, Y. (2020). Draw with me: human-in-the-loop for image restoration. In Proceedings of the 25th International Conference on Intelligent User Interfaces (pp. 243–253).

2.6
Automated Program Synthesis

Authors:
Dr. Harald Rueß, Dr. Xin Ye

Program synthesis is the task of constructing a program with a problem description as input and an output program for solving the given problem description. This is a long-standing challenge in the field of AI, and program synthesis sub-sumes a number of prominent AI techniques such as AI planning and reinforcement learning. Program synthesis is also a fundamental tool in cognitive sciences for understanding human cognitive capabilities such as concept learning and social reasoning. From a software engineering perspective, program synthesis fundamentally changes our way of programming in that we specify "what" should be achieved instead of explicitly programming "how" problems are being solved.

The proposed benefits of program synthesis for AI engineering are manifold. First of all, programmers of AI and any other software system are being freed from the Turing tarpit of painstakingly sequencing program statements. Instead, resource-efficient solutions are automatically generated, possibly based on libraries for encoding programming knowledge such as algorithmic theories and corresponding optimization strategies. Second, program synthesizers may produce correct-by-construction and super-optimized programs, which do not need to be verified or tested. Third, programs may be automatically re-synthesized and self-adapted to reflect ever-evolving problem descriptions, system capabilities, available resources, operating environments, and extra-functional requirements (dependability, robustness, trustworthiness). Fourth, program synthesis realizes the main step in synthesizing solution strategies for goal-oriented cognitive systems. And finally, program synthesis is used for automatically synthesizing AI system runtime monitors. In this way, program synthesis is a key element towards meaningful control for AI-based mission-critical systems, whereby learning-enabled and largely untrusted AI components are being monitored for consistency, confidence and violations of safety constraints.

Program synthesis may in the near future also be used in a more active, even interactive, manner by solving uncertain and partially specified problems or partially known users' intent by cautiously probing these situations, for example, by minimizing surprises and associated free energy, until sufficiently trustworthy solution strategies and programs can be generated. For example, such an AI-based agent may actively trigger focused perception actions for strengthening its certitude that, say, the traffic light ahead, which may be partially occluded, indeed is green.

There is a plethora of variations on program synthesis, depending on the class of programs to be synthesized (sequential, reactive, probabilistic), the means of problem specification and applicable regulations (natural language, logical expressions, examples), and the nature of the underlying search for solution programs (enumerative and stochastic search, constraint solving, reinforcement learning). In the context of specific applications and scenarios, one is interested in properties of the synthesized programs including correctness (hard/soft) and performance (resource-efficency), dependability with respect to internal and external defects, and generalizability/robustness with respect to both uncertain ignorance and knowledge.

Whereas we are still a far cry away from synthesizing general AI problems,[39] there has been substantial progress in program synthesis with lots of interesting practical applications. Hereby, it is important to concentrate on well-defined and restricted classes of problem descriptions with corresponding domain- and problem-specific programming languages.

In particular, efficient constraint-based engines such as *satisfiability modulo theories* (SMT) form the computational backend of many deductive program synthesis approaches. There is also deep and productive connections of program synthesis with machine learning and inductive programming. The goal of inductive programming is to generate a function that matches a given set of input-output examples. Indeed, the approximation of functions based on input-output examples and the "learning" of corresponding artificial neural network structures is a special case of program synthesis, whereby the searchable program space is limited to neural network structures that has its own specialized set of algorithms for deriving a function that matches a dataset . By contrast, program synthesis focuses on general algorithms that can work with more general classes of programs. Machine learning (neural synthesis, transfer learning) techniques have proven almost indispensable for guiding the search for suitable programs in program synthesis based on prior experience.

For the importance of program synthesis in engineering robust and trustworthy AI systems, at fortiss we developed a long-standing and continuing research line on the automated generation of, mostly, embedded control programs. These kinds of programs are usually reactive; that is, they are continually reading inputs (such as from sensors) and computing corresponding outputs. Thus, reactive programs are the underlying model of the sense-compute-act triad of cognitive cyber physical systems and Internet of Things applications, which are increasingly acting autonomously.

So far we have mainly concentrated on the two complementary challenges of synthesizing reactive programs both in-the-small and in-the-large.

- Reactive program synthesis
  Synthesizing a reactive program given its specification in a temporal logic formalism
- Coordination program synthesis
  Synthesizing a reactive coordination program between a given set of reactive and interacting programs to solve a given problem description as specified in temporal logic

These and many other program synthesis problems are reduced to solving exists-forall quantified logic constraints. Indeed, we have been developing the ∃∀SMT constraint engine for solving these kinds of constraints based on a game-like coupling of an exists-SMT solver with a forall-SMT solver and the mutual exchange of generated knowledge (Cheng et al, 2014b). An alternative approach to strategy synthesis is based on computing winning strategies in mu-calculus by

[39] In particular, solving the program synthesis problem by program synthesis itself, and so on...

means of direct evaluation of fixed points and partial winning strategies (Hofmann et al, 2016); notice that common program specification logics such as CTL or LTL are included in mu-calculus.

Reactive program synthesis

Embedded control software in the manufacturing and processing industries is usually developed using specialized programming languages such as ladder diagrams or other IEC 61131-3 defined languages. Programming in these rather low-level languages is not only error-prone but also time- and resource-intensive. On the other hand, these control programs are often used in mission- and even safety-critical scenarios, thereby placing increased and demonstratable demand on correctness, dependability and safety.

In the community of robust controller synthesis, assumption-guarantee (AG) specification is used to describe the behaviors of environment and systems[40]. That is, the specification is of the form A->G where A is an environment assumption and G is a guarantee. If the environment satisfies the assumption, the system reacts correctly as intended. Generalized reactivity(1) (GR 1) is one strict AG specification for reactive synthesis problems[41] and has been used in various applications such as, robotics, scenario-based specifications, aspect languages and event-based behavior. Counter-guided strategy is an approach for correcting an unrealizable specification such as correction of GR 1 specification by adding assumptions on the environment[42].

At fortiss we have been developing an actor-based algorithm for synthesizing reactive embedded programs (Cheng et al, 2017b; Cheng et al,2016), which is also useful for demonstrating safety. This novel class of synthesis algorithm generates, for a given specification in a suitable subset of linear temporal logic (LTL) called GXW, a structured dataflow program by adequately wiring and instantiating pre-specified compute actors. Actor-based synthesis for GXW specifications is in PSPACE compared to 2EXPTIME-completeness of full-fledged LTL synthesis. Under some further reasonable syntactic restrictions on the GXW fragment actor-based synthesis can even be shown to be in coNP. The biggest distinction between GXW and GR1 is that, in GXW, traceability requirements can be supported via operators G, X^i and W where G is the universal path quantifier, X^i abbreviates i consecutive next steps and W is the weak until operator while GR 1 only handles specifications involving assertions over initial states, safety constraints relating the current and next state, and goals for liveness properties. For example, there is a requirement in an automatic door open close system (that cannot be expressed in GR1 formula):

“When someone enters the infrared sensing field, opening motor starts working to open the door automatically until the door touches the opening limit switch.”

This specification can be described as a GXW formula: $G((\neg in0 \wedge X\ in0)^1 \rightarrow X(out0\ W\ in2))$ where $in0$ is true when someone enters the sensing field, $out0$ denotes the opening motor and $in2$ denotes an opening limit switch.

For each GXW formula, our algorithm constructs actors and wirings for monitoring low-level events by mimicking the DNF formula structure. As an actor defines a Mealy Machine corresponding to one GXW formula, the main advantage of actor-based synthesis compared to earlier automata-based approaches

[40] Jonsson, Bengt, and Tsay Yih-Kuen. “Assumption/guarantee specifications in linear-time temporal logic.” Theoretical Computer Science 167.1-2 (1996): 47–72.

[41] Bloem, Roderick, et al. “Synthesis of reactive (1) designs.” Journal of Computer and System Sciences 78.3 (2012):

[42] Alur, Rajeev, Salar Moarref, and Ufuk Topcu. “Counter-strategy guided refinement of GR (1) temporal logic specifications.” Formal Methods in Computer-Aided Design. IEEE, 2013. 911–938.

to LTL synthesis is that actor-based synthesis maintains the traceability between individual requirements and the generated controller code blocks. Indeed, such a line-by-line tracing is required by most safety engineering and certification standards. The structured approach of actor-based synthesis also forms the basis of automated incremental change.

Indeed our experimental results suggest that GXW is sufficiently expressive. Also, GXW synthesis scales well to synthesis problems with 20 input and output ports and beyond, which seems to be sufficient for most control problems as encountered in industrial practice. In fact, we have successfully applied actor-based synthesis to more than 70 embedded control scenarios from industrial practice (PLC control of wind mills) and industrial training cases (CODESYS 3.0, AC500). Other applications of actor-based synthesis include the interactive analysis of requirements for embedded control applications (Lúcio et al, 2017a; Lúcio et al, 2017b) as expressed in the industrial EARS requirement specification language. The automated generation of industrial-scale PCS programs based on game solving is demonstrated in Cheng et al, 2014a. Further applications such as game-based production in Industry 4.0 scenarios, where production is modeled as a game between the production facility and the workpieces to be produced (Cheng et al, 2013b; Cheng et al, 2012), may also possibly be expressed in terms of actor-based reactive synthesis.

Figure 10.
Selected requirements of computer-aided resuscitation.

Req-08	If Air Ok signal remains low, auto-control mode is terminated within 3 seconds.
Req-17	When auto-control mode is entered, eventually the cuff will be inflated.
Req-28	If a valid pressure is unavailable within 180 seconds, manual mode should be triggered.
Req-32	If pulse wave or arterial line is available, and cuff is selected, corroboration is triggered.
Req-42	When auto-control mode is running, and the arterial line or pulse wave or cuff is lost, an alarm should sound within 60 seconds.
Req-44	If pulse wave and arterial line are unavailable, and cuff is selected, and blood pressure is not valid, next manual mode is started.

Extensions to actor-based synthesis include increased expressivity of GXW by means of numerical constraints (Cheng et al, 2013a). A semantics-driven (cmp. ARSENAL) translation of natural language specifications into GXW formulas is presented in Yan et al, 2015, and forms the basis for formal consistency checks of natural language specifications. These consistency checks together with the realization of natural language specifications by means of actor-based synthesis open up the possibility of directing computer systems by means of everyday natural language, whereby the computer system itself is used both as a workhorse for program generation and a critical companion for validating and improving specifications. Such a critical programming companion may, for example, query programmers or suggest improvements based on identified ambiguities, imprecision, underspecifications, inconsistencies, or potential safety violations, and more generally, also for probing the programmers' intent.

Programming from natural language instructions (see Figure 10) has been successfully demonstrated in Yan et al, 2015 by means of the computer-aided resuscitation algorithm CARA, which monitors and controls the operation of an infusion pump for driving resuscitating fluids into a patient's blood stream. Clearly, CARA is highly safety-critical.

Currently we are extending actor-based synthesis for automatically generating reactive controllers with real-time constraints (Ye et al, 2021), for controllers with analog input and output signals, and for synthesizing robust controllers, which may tolerate small deviations based on, say, sensor noise, sensor failure, or even sensor attacks. This novel approach of synthesis algorithm generates from templates given in a timed GXW specification, an extension of GXW with metric intervals. With Timed GXW, it can be described as a formula via duration timed operator G_I, W_I, X_I, where I is time interval. The generated controller is actor-based using timed synchronous dataflow without circles. An actor defines an event-clock automaton so that time interval can be handled as transition execution time of the automaton. First, our algorithm prepares I/O ports and creates high-level controllers based on timed GXW pattern. Then, for each sub-formula, actors and wirings for monitoring low-level events can be constructed. Finally, SMT satisfiability checking is applied to guarantee nonexistence of potential conflicts between different formulas.

For example (in Figure 11), we specify and apply actor-based synthesis for a controller automatically infusing the container with liquids A and B in order when START is pressed (in0 is true). Inputs and outputs are as follows:

A Input in0 is true when START is pressed and out0 will be true (the valve is opened for infusing liquid A) until the level reaches the low-level float sensor (specification S1);

B input int1 is true when the level reaches the low-level float sensor and output out1 will be true (the valve is opened for infusing liquid B) until the level reaches the high-level float sensor (specification S2);

C input in2 is true when the level reaches the high-level float sensor and output out3 will be true activating the agitator for 60 sec (specification S3). Also, output t_{start} models the triggering of a 60-sec timer (specification S4);

D input $t_{1expire}$ is true when the 60-sec timer expires and output out3 is false (the agitator motor stops working, specification S5). Also, output2 is true and the mixture will drain out of the container (specification S6).

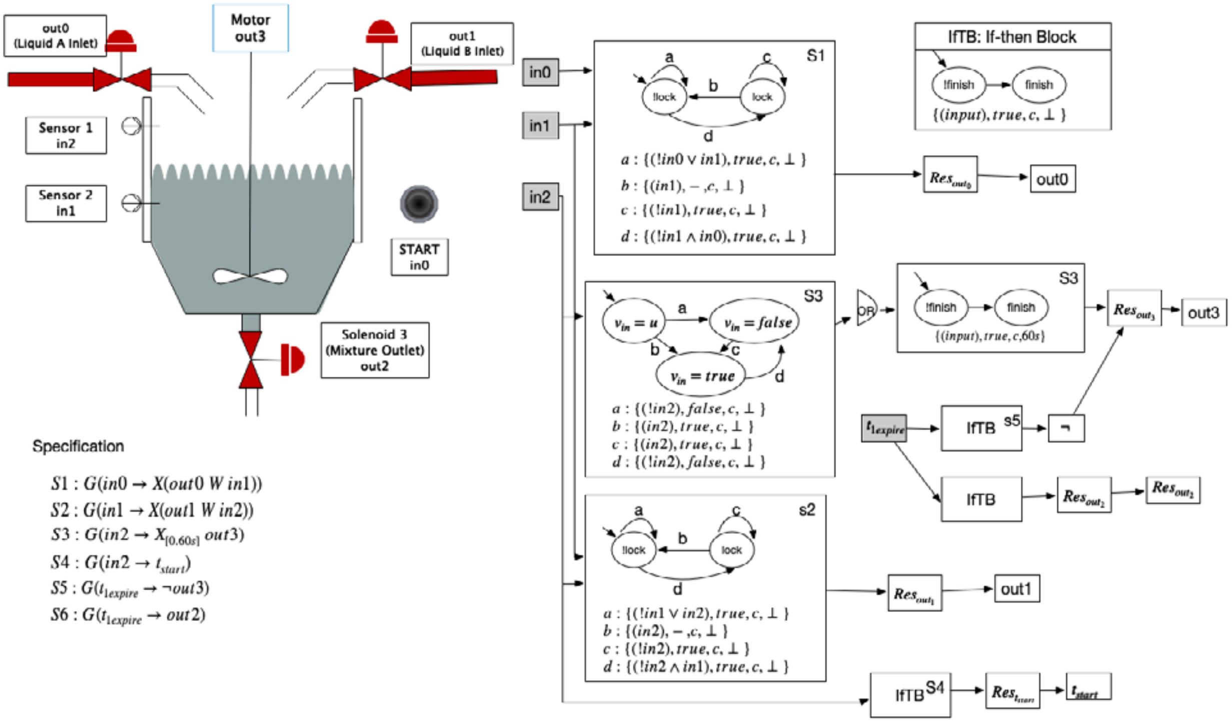

out0 (Liquid A Inlet)
Motor out3
out1 (Liquid B Inlet)
Sensor 1 in2
Sensor 2 In1
START in0
Solenoid 3 (Mixture Outlet) out2
Specification
$S1 : G(in0 \rightarrow X(out0\ W\ in1))$
$S2 : G(in1 \rightarrow X(out1\ W\ in2))$
$S3 : G(in2 \rightarrow X_{[0.60s]}\ out3)$
$S4 : G(in2 \rightarrow t_{start})$
$S5 : G(t_{1expire} \rightarrow \neg out3)$
$S6 : G(t_{1expire} \rightarrow out2)$
in0
in1
in2
S1
!lock
lock
a
b
c
d
$a : \{(!in0 \vee in1), true, c, \perp\}$
$b : \{(in1), -, c, \perp\}$
$c : \{(!in1), true, c, \perp\}$
$d : \{(!in1 \wedge in0), true, c, \perp\}$
IfTB: If-then Block
!finish
finish
$\{(input), true, c, \perp\}$
Res_{out_0}
out0
S3
$v_{in} = u$
$v_{in} = false$
$v_{in} = true$
$a : \{(!in2), false, c, \perp\}$
$b : \{(in2), true, c, \perp\}$
$c : \{(in2), true, c, \perp\}$
$d : \{(!in2), false, c, \perp\}$
OR
S3
!finish
finish
$\{(input), true, c, 60s\}$
Res_{out_3}
out3
$t_{1expire}$
IfTB s5
¬
IfTB
Res_{out_2}
Res_{out_2}
s2
!lock
lock
$a : \{(!in1 \vee in2), true, c, \perp\}$
$b : \{(in2), -, c, \perp\}$
$c : \{(!in2), true, c, \perp\}$
$d : \{(!in2 \wedge in1), true, c, \perp\}$
Res_{out_1}
out1
IfTB S4
$Res_{t_{start}}$
t_{start}

Figure 11.
Actor-based controller realizing infusion liquids

The corresponding actor-based controller shown in the right part of Figure 11 with the given timed GXW specification is made explicit by superscripting actors with index (i) i.e. the actor has been introduced due to the i-th specification.

In this work, transferability and, more generally, robustness with respect to different operating environments may be obtained, for example, by constructing the potentially weakest constraints on these operating environments while still being able to generate realizable solution programs. Online versions of algorithms for reactive program synthesis may also be used for realizing the cognitive capabilities of AI-based agents in the future. These kinds of applications may require real-time or even any-time response, at the possible expense of other requirements such as efficiency, correctness, or generality of the solution program.

Coordination program synthesis

Coordination problems arise naturally in many CPS/IoT settings. In a so-called smart building, various sensors, heating and cooling devices must work in concert to maintain comfortable conditions. In a fully automated factory, robots with specialized capabilities must collaborate to carry out manufacturing tasks. Typically, the individual agents are reactive and a centralized coordinator provides the necessary overall guidance to carry out a task so that the combined system satisfies the specification of its desired behaviors such as the users' intent.

A coordination program must work in the presence of several complicating factors such as concurrency, asynchrony, and distribution and should recover gracefully from agent failures and handle noisy sensor data. All this complicates the design of coordination programs. It is often the case, however, that the task itself can be specified easily and compactly.

We therefore consider whether it is possible to automatically synthesize a reactive coordination program from a description of a (finite) set of interacting reactive programs and a safety specification (mutual exclusion, free of deadlocks, fault repair, conflict resolution). Interface descriptions of the individual programs in Java may, for example, be provided in terms of behavioral types in OSGi (Blech et al, 2012). There is a possible non-determinism in that several different interactions may be selected for execution at any point in time. Now, priority synthesis ensures given safety specifications by restricting the set of possible behaviors by means of generating a suitable, and possibly minimal, set of priorities a < b, since such a priority always prefers the execution of interaction a over b (Cheng et al, 2011b; Cheng et al, 2011c).

Priorities have been shown to be an expressive concept for coordinating interacting programs in embedded and autonomous systems, even though they are strictly less expressive than automata-based coordination programs in the Ramadge-Wonham controller synthesis framework. It is easy to see that priority synthesis for simple safety specifications is NP-complete (Cheng et al, 2011d). Also, priorities, for their stateless nature, facilitate distributed control programs (Cheng et al, 2011a). The extension of stateless priorities to state-dependent priorities is investigated in Herrera, 2020; Herrera et al, 2020, as stateless priorities may, at times, be overly restrictive.

Our main case studies for priority synthesis include scheduling in multicore processors for 3D image processing applications (Cheng et al, 2011a), and the synthesis of safe coordinating priorities of the DALA robot from LAAS with a software control stack of around 170k lines of code.[43] The crash-free scheduling of transportation robots in a factory setting, and the synthesis of stateful priorities for establishing a collision-free CSMA/CD network are investigated in Herrera, 2020; Herrera et al, 2020.

Finally, we consider the complementary technique of parameter synthesis for parametric timed programs and their coordination. Individual machines in flexible production lines, for example, explicitly expose capabilities at their interfaces by means of parametric skills such as drilling for instance. Given such a set of configurable machines, a line integrator is faced with the problem of finding and tuning parameters for each machine such that the overall production line implements given safety and temporal requirements in an optimized and robust fashion. In Cheng et al, 2016b we formalize these kinds of problems as parameter synthesis problems for systems of parametric timed automata, where interactions are based on skills. Parameter synthesis problems for interaction-level LTL properties are then translated to parameter synthesis problems for state-based safety properties. For safety properties, synthesis problems are solved by checking the satisfiability of ∃∀SMT constraints. The feasibility of this approach is demonstrated in Cheng et al, 2016b by solving typical machine configuration problems as encountered in industrial automation. Finally, compositional parameter synthesis for parametric timed systems is studied in Aştefănoaei et al, 2016.

Software

- Autocode4 synthesizes synchronous dataflow controllers from the GXW subset of linear temporal logic specifications. This intermediate format may be translated to, among others, Lustre/Scade, LabView, and Ptolemy II,

[43] Hereby, we have been relying on interface descriptions by Verimag on the individual control software components.

Matlab Simulink, and IEC 61131-3 continuous function charts. It is based on the actor-based approach to reactive synthesis as developed in Cheng et al, 2017b; Cheng et al, 2016a. Autocode4 also supports interactive environment specification based on instantiating specifications patterns and by analyzing causes of unrealizability. Autocode4 is available under the LGPL 3.0 license at http://autocode4.sourceforge.net.

- CrESto coordinates the actions of a group of other reactive programs so that the combined system satisfies a given safety specification. It is based on the algorithm for synthesizing transition priorities as developed in Herrera et al, 2020. CrESto is able to obtain stateful priorities that avoid reaching error states in several real-world examples. An extension to the query language of CrESto supports queries with data variables that frees users from modeling networks and queries just for querying data values and allows users to design more natural networks and queries.

LITERATURE

Aştefănoaei, L., Bensalem, S., Bozga, M., Cheng, C. H., & Rueß, H. (2016). Compositional parameter synthesis. In International Symposium on Formal Methods (pp. 60–68). Springer, Cham.

Blech, J. O., Falcone, Y., Rueß, H., & Schätz, B. (2012). Behavioral specification based runtime monitors for OSGi services. In International Symposium On Leveraging Applications of Formal Methods, Verification and Validation (pp. 405–419). Springer, Berlin, Heidelberg.

Cheng, C. H., Lee, E. A., & Rueß, H. (2017b). autoCode4: Structural controller synthesis. In International Conference on Tools and Algorithms for the Construction and Analysis of Systems (pp. 398–404). Springer, Berlin, Heidelberg.

Cheng, C. H., Hamza, Y., & Rueß, H. (2016a). Structural synthesis for GXW specifications. In International Conference on Computer Aided Verification (pp. 95–117). Springer, Cham.

Cheng, CH., Astefaneoaei, L., Rueß, H., Rayana, S., Bensalem, S. (2016b). Timed Orchestration for Component-Based Systems, arXiv: 1504.05513, 2016.

Cheng, C. H., Huang, C. H., Rueß, H., & Stattelmann, S. (2014a). G4LTL-ST: Automatic generation of PLC programs. In International Conference on Computer Aided Verification (pp. 541–549). Springer, Cham.

Cheng, C. H., Bensalem, S., Rueß, H., Shankar, N., & Tiwari, A. (2014b). Efsmt: A logical framework for the design of cyber-physical systems. Cyber-Physical System Architectures and Design Methodologies (CPSArch).

Cheng, CH., Lee, E.(2013a) Numerical LTL Synthesis for Cyber-Physical Systems, work-in-progress report, ARXiv: 1307.3722, 2013.

Cheng, C. H., Geisinger, M., & Buckl, C. (2013b). Synthesizing controllers for automation tasks with performance guarantees. In International SPIN Workshop on Model Checking of Software (pp. 154–159). Springer, Berlin, Heidelberg.

Cheng, C. H., Geisinger, M., Rueß, H., Buckl, C., & Knoll, A. (2012). Game solving for industrial automation and control. In 2012 IEEE International Conference on Robotics and Automation (pp. 4367–4372). IEEE.

Cheng, C. H., Bensalem, S., Yan, R., Rueß, H., Buckl, C., & Knoll, A. (2011a). Distributed Priority Synthesis and its Applications. arXiv preprint arXiv:1112.1783.

Cheng, C. H., Bensalem, S., Chen, Y. F., Yan, R., Jobstmann, B., Rueß, H., ... & Knoll, A. (2011b). Algorithms for synthesizing priorities in component-based systems. In International Symposium on Automated Technology for Verification and Analysis (pp. 150–167). Springer, Berlin, Heidelberg.

Cheng, C. H., Bensalem, S., Jobstmann, B., Yan, R., Knoll, A., & Rueß, H. (2011c). Model construction and priority synthesis for simple interaction systems. In NASA Formal Methods Symposium (pp. 466-471). Springer, Berlin, Heidelberg.

Cheng, C. H., Jobstmann, B., Buckl, C., & Knoll, A. (2011d). On the hardness of priority synthesis. In International Conference on Implementation and Application of Automata (pp. 110–117). Springer, Berlin, Heidelberg.

Cheng, C. H., Rueß, H., Knoll, A., & Buckl, C. (2011e). Synthesis of fault-tolerant embedded systems using games: From theory to practice. In International Workshop on Verification, Model Checking, and Abstract Interpretation (pp. 118–133). Springer, Berlin, Heidelberg.

Herrera, C. (2020). Stateful Priorities for Precise Restriction of System Behavior. In 2020 16th European Dependable Computing Conference (EDCC) (pp. 69–76). IEEE.

Herrera, C., Cruz, N., & Quintero, R. (2020). CrEStO: A Tool for Synthesizing Stateful Priorities. In 2020 16th European Dependable Computing Conference (EDCC) (pp. 143–146). IEEE.

Lúcio, L., Rahman, S., Cheng, C. H., & Mavin, A. (2017a). Just formal enough? automated analysis of ears requirements. In NASA Formal Methods Symposium (pp. 427–434). Springer, Cham.

Lúcio, L., Rahman, S., bin Abid, S., & Mavin, A. (2017b). EARS-CTRL: Generating Controllers for Dummies. In MODELS (Satellite Events).

Hofmann, M., Neukirchen, C., & Rueß, H. (2016). Certification for mu-calculus with winning strategies. , In: Proc. of the International Symposium on Model Checking Software, pp. 111–128, LNCS, vol. 9641, Springer, 2016.

Yan, R., Cheng, C.H. & Chai, Y.(2015). Formal consistency checking over specifications in natural languages. In 2015 Design, Automation & Test in Euro.

Ye, X., Rueß, H. (2021). Timed Actor-Based Synthesis, work in progress, 2021

2.7 Edge AI

Author:
Prof. Dr. Rute Sofia

Rationale

In the quest to assist smart data computation in a *Next Generation Internet of Things*, research trends concern a decentralisation of Internet services and of computational as well as network architectures. Such trends attempt to best serve the mobility of devices and users, the need for data and user privacy, the larger volumes of sensitive data to be analyzed, and the requirements to handle such data. This is giving rise to alternative ways to provide data exchange in *Inter-net of Things (IoT)* environments, as occurs for instance, with the paradigms of edge/fog computing[44]. By definition, edge/fog computing envisions a smooth migration of applications and services between different physical and virtual machines to best meet the application requirements. In practice, such migration still requires a high degree of human intervention, as can be observed in the ETSI *mobile edge computing* (MEC) architecture[45], where supported scenarios consider migration mostly for the purpose of backup and restore of applications, or for redundancy.

A Next Generation IoT will, however, have to handle mobility both in regards to physical and virtual machines, as well as in regards to data sources (*traffic and data locality*). It has also to handle applications across edge and cloud networks, in a way that is not necessarily tied to network policies or geographical boundaries[46]. Today, the definition of *edge* is elastic and not tied to a specific infrastructure boundary. As a result, the edge component is also reaching end-user devices, such as smartphones, or smart sensors placed in industrial environments as represented in Figure 12.

To best support next generation IoT applications such as augmented reality (AR) it is necessary to integrate intelligence into the edge network such as AI methods for training and behavior inference. This needs to be accommodated both at an individual level (within one single device or cyber physical system) and at a collective perspective (a set of autonomous, smart devices, cooperating to best automate data exchange). Intelligence at the edge, also known as edge AI, implies concepts where even the smallest devices and machines around us are able to sense, learn from, and respond to their surroundings. This enables machines in a public space or in a factory, for instance, to make higher-level decisions, act autonomously, and report back relevant errors or improvements to the user or the cloud. Reactions (inference) can then be sent to the cloud, or be used for some physical actuation in the local environment. The captured data can be stored in a decentralized way across different edge networks and sent to the cloud for specific processing derived from behavioral learning and inference. Pre-training and learning are traditionally part of a continuous process, so that edge devices can learn in close-to-real-time, while they process captured information.

[44] Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. Fog computing and its role in the internet of things. In Proceedings of the first edition of the MCC workshop on Mobile cloud computing (pp. 13–16). 2012.

[45] Giust, F., Costa-Perez, X., & Reznik, A. (2017). Multi-access edge computing: An overview of ETSI MEC ISG. IEEE 5G Tech Focus, 1(4), 4.

[46] Networld, " Smart Networks in the context of NGI" Strategic Research and Innovation Agenda 2021 – 27. September 2020.

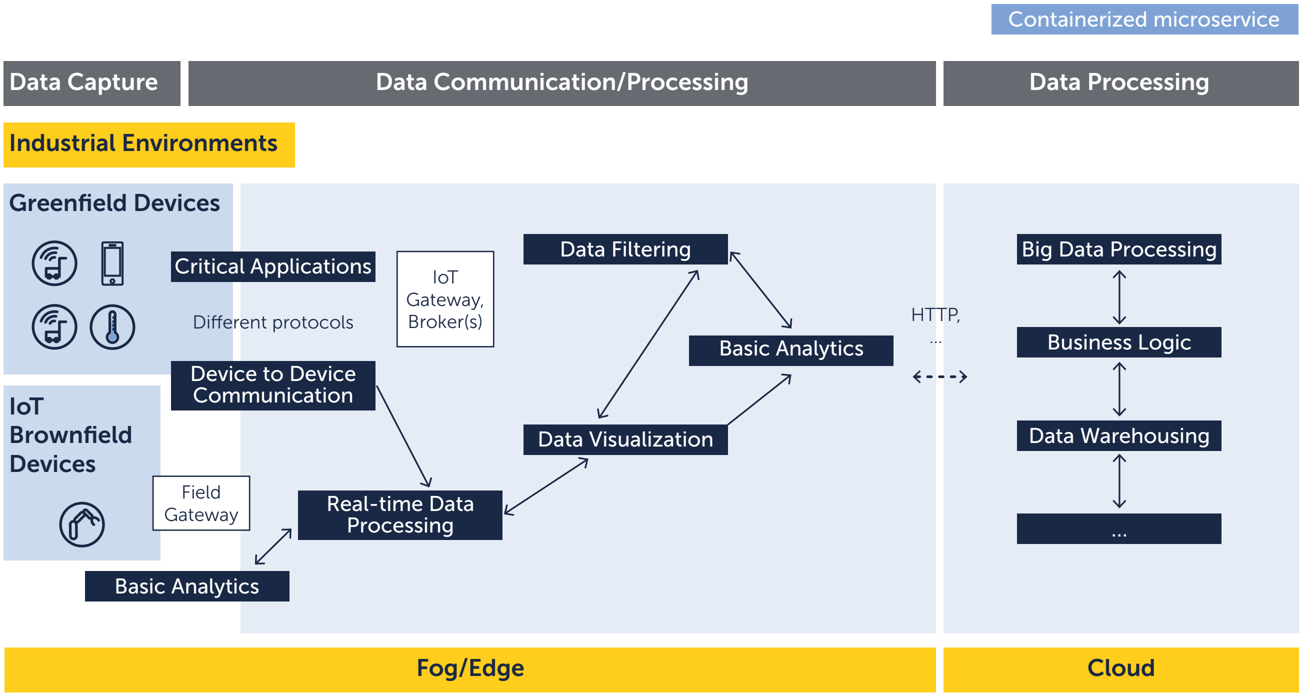


Figure 12. Cloud, fog/edge in the Industrial IoT context

This corresponds to the traditional architectural model that supports intelligence (AI training and classification) across edge/cloud environments.

However, the number of available data sources in different environments continues to increase. In 2021, 850 ZB of data are expected to be collected from devices such as machines, sensors and personal smart devices (people)[47]. Moreover, sensors such as accelerometers, GPS, microphones or cameras in personal mobile devices provide the ability to leverage new types of data, referred to as "smart data" or "small data", resulting from tracking various aspects of peoples' routines, such as roaming habits, application usage or location preferences. Small data brings in a new level of granularity in terms of features, and corresponds also to lower volumes of data than „big data", which introduces new problems in terms of data validation and processing.

The integration of intelligence into the edge network, such as training and classification tasks, is thus a key aspect to achieving service decentralization and a much desired aspect in IoT environments, especially industrial. Still, the majority of today's research is focused on bringing intelligence to the so-called "*near edge*" infrastructure, for which a reference architecture is the ETSI MEC, where powerful computational devices are placed in an area still within reach of the operator but closer to data sources, thus often simply replicating, at a lesser extent and for a specific local purpose, the cloud computational environment. This does not suffice to support decentralized services in an IoT. It is necessary to support intelligence in "*far edge*" scenarios, where the "far edge" corresponds to the infrastructure deployed within the customer premises closer to data sources, such as a production environment, shopping mall, stadium or a home. Bringing intelligence to the far edge, in a way that is relevant to further advancing next generation IoT applications, is the main aspect under development by fortiss in the context of decentralized edge computing (see Figure 13).

[47] Networld2020 "SatCom Resources for Smart and Sustainable Networks and Services" Nov. 2019

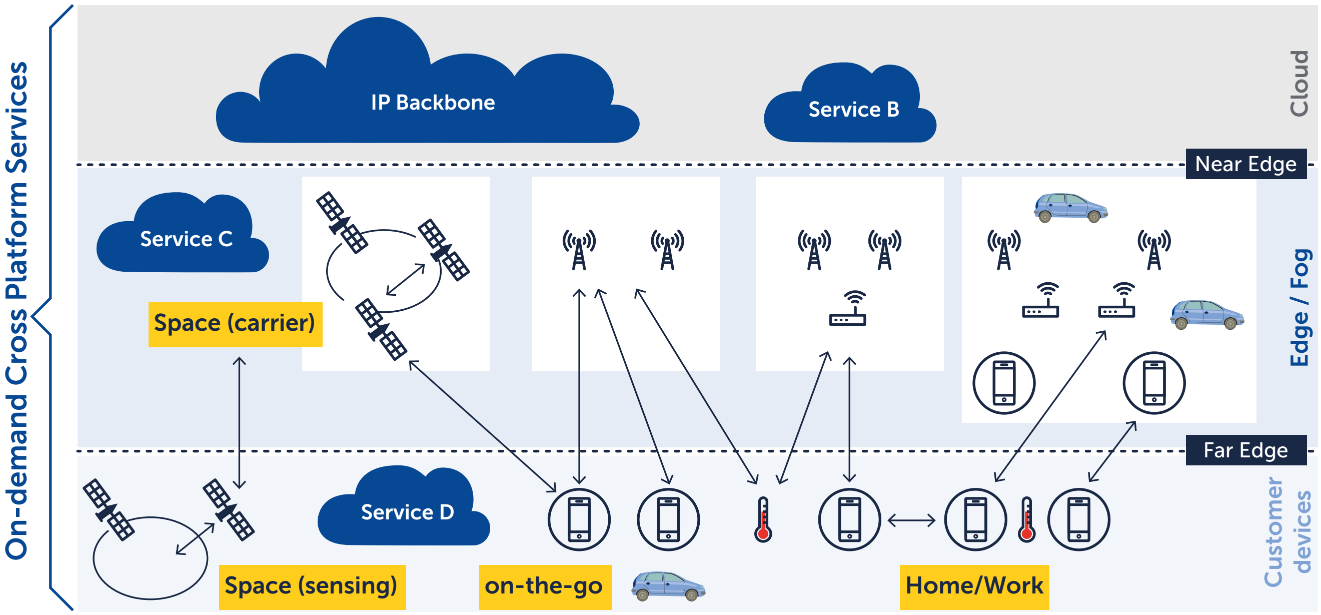


Figure 13.
Decentralized edge computing, far and near edge representation. decentralization requires intelligence at the edge.

Scope/our approach

Bringing intelligence to the far edge network requires devising decentralized edge computing architectures, beyond the MEC architecture, and understanding 1) how to support AI engineering on such decentralized scenarios, via distributed AI approaches 2) which methods best serve the challenges and constraints posed by realistic far edge scenarios, in particular in regards to Industrial IoT scenarios 3) how can ML models be adapted to serve the constraints of IoT sensors (embedded devices), supporting as well challenges such as intermittent connectivity; mobility management.

In regards to the operationalization of AI in decentralized edge scenarios, distributed AI methods, such as federated learning, are starting points for the support of intelligence in the edge network. The use of distributed AI methods nonetheless needs to consider new challenges such as the constraints of different devices, not forgetting personal smart devices (such as smartphones, which today are the basis for *mobile crowd sensing* services) and yet, at the same time, considering new frontiers such as smart satellite constellations. Furthermore, such discussions must not simply consider individual devices as the basis, but also how to optimally provide the underlying networking architecture to best support distributed model training and eventually classification. For instance, it is important to support design aspects such as mobility management and privacy/accountability. Therefore, to better support these dynamic environments, several steps are critical to achieve a better edge cloud continuum: (1) service selection and adaptation (2) dynamic computation offloading (3) embedded AI performance evaluation.

Research under development

Intelligent edge solutions must be able to handle higher levels of automation, mobility in terms of both physical and virtual machines, and data sources (traffic

[48] Yang, Qiang, et al. "Federated machine learning: Concept and applications." ACM Transactions on Intelligent Systems and Technology (TIST) 10.2 (2019): 1–19.

[49] Yu, Tao, Yue Zhang, and Kwei-Jay Lin. "Efficient algorithms for Web services selection with end-to-end QoS constraints." ACM Transactions on the Web (TWEB), 2007.

→

and data localization). In addition, future applications will also need to take into account that containerized applications run across edge and cloud networks in a way that is not necessarily bound by network policies or geographic boundaries, but in contrast to the context of the different stakeholders[48]. The edge definition itself is elastic, and therefore it is assumed that an edge node can be part of different embedded devices. The edge node is also expected to be integrated with personal end-user devices or smart sensors, which increases the need to incorporate mechanisms that can cope with variability in resources, location and data sets/data types.

Service selection and adaptation

Related work has proposed service selection and adaptation via methods such as specific heuristics[49] or the application of genetic algorithms[50] that aim to find near-optimal compositions, such as compositions respecting overall *quality of service (QoS)* and *quality of experience (QoE)* constraints, while maximizing a QoS/QoE utility function. The composition of services in current cloud-edge big data/AI applications such as for smart industry and IoT usually follows a pipeline pattern in which stream and batch data (potentially recorded at the edge) is processed by multiple services/tasks in order to derive the desired results. This pattern has been formalized and implemented in products such as Google's data flow[51]. These new pipelines add new requirements and challenges to service selection and adaptation as they inherently contain complex trade offs between computation and performance (resulting accuracy) and errors introduced in early components cascade through the overall pipeline, affecting overall performance and making it impossible to treat the problem as an independent selection and adaptation of services.

Initial approaches address this problem with reasoning across the pipeline components in a probabilistically manner, allowing the user to manually decide the adequate trade-off[52]. Recently, *reinforcement learning (RL)* [53] has been successfully applied to device selection for execution[54] as well as optimization of overall pipelines using among others, meta-reasoning techniques to ensure an overall optimization of the pipeline[55] [56] [57]. The current research under development in fortiss is expected to advance recent progresses in making RL applicable for distributed system optimization by combining data-driven knowledge into a novel RL approach. Specifically, RL will be combined with guiding and constraint functions to ensure an accelerated warm-up time of the RL agent in live-systems and to avoid undesired actions in unsafe system states that are otherwise common in the exploration phase of traditional RL. The knowledge of single RL agents will be abstracted and shared with other agents through an adaption of deep multitask RL[58].

Dynamic offloading

Next generation IoT applications rely on a *microservice architecture* model[59]. Microservice architectures support applications as a set of fine-grained services, loosely coupled, interacting via lightweight protocols. The deployment of microservices is supported by virtualization technology, in which container runtime technology, such as Docker[60] [61], is becoming the most popular solution. The reason for the increasing adhesion to container solutions, in particular *container*

[50] Canfora, Gerardo, Massimiliano Di Penta, Raffaele Esposito, and Maria Luisa Villani. "An approach for QoS-aware service composition based on genetic algorithms." In Proceedings of the 7th annual conference on Genetic and evolutionary computation, pp. 1069–1075. 2005.

[51] Google Data Flow. Available on https://cloud.google.com/dataflow. Consulted on 10.04.2020.

[52] Raman, Karthik, Adith Swaminathan, Johannes Gehrke, and Thorsten Joachims. "Beyond myopic inference in big data pipelines." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 86–94. 2013.

[53] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[54] Mirhoseini, Azalia, Hieu Pham, Quoc V. Le, Benoit Steiner, Rasmus Larsen, Yuefeng Zhou, Naveen Kumar, Mohammad Norouzi, Samy Bengio, and Jeff Dean. "Device placement optimization with reinforcement learning." In Proceedings of the 34th International Conference on Machine Learning, 2017.

[55] Modi, Aditya, Debadeepta Dey, Alekh Agarwal, Adith Swaminathan, Besmira Nushi, Sean Andrist, and Eric Horvitz. "Metareasoning in Modular Software Systems: On-the-Fly Configuration using Reinforcement Learning with Rich Contextual Representations." arXiv preprint arXiv:1905.05179. 2019.

as a service (CaaS) for which there is an expected *compound annual growth rate (CAGR)* of 34.4% until 2025[62], concerns intrinsic features such as isolation and portability, as well as increased security due to isolation of the code being run (isolated user space in a physical or virtual machine).

While CaaS platforms are today more common as a service model for cloud providers such as Google, Microsoft and IBM, this type of platform brings in advantages for a seamless cloud/edge operation interconnecting *operational technology (OT)* and *information technology (IT)* environments. For instance, specific microservices can be pushed into the edge network, improving operational aspects such as energy consumption, security or reliability. In an exemplar deployment scenario, specific components of an application are distributed and engineered to reach a desired level of QoE. To reduce operational costs, all these factors need to be taken into account during the application design phase. Therefore, often the underlying CaaS architecture is designed to serve a specific and initial application model, not taking into consideration the changing application requirements over time, the context (a service being offered in a mobile infrastructure), or even the perceived QoE. Moreover, the CaaS management, performed via *container orchestration*[63] tools such as Kubernetes[64], Docker Swarm[65] or Apache Mesos[66], assists in deciding "how" and "where" to run application workloads and how to configure the required (overlay) infrastructure that interconnects, via TCP/IP, the different physical and virtual machines. Aspects that are taken into consideration by container orchestrators, concern a semi-automated way of deploying, scaling and managing containerized applications. Tools such as Kubernetes, the de facto container orchestration which has a cloud market adoption rate of 86% [67], provide the means to manage containerized applications across cloud/edge environments. Such tools provide scripting and user interfacing that supports basic system configuration to setup *clusters* [68] of containers, their processes (*pods*), the required interconnection for data exchange and discovery, in the form of a network overlay built on TCP/IP. Typically, clusters comprise containers in the cloud and in the edge network. However, continuous cloud/edge support is still not feasible today, as container orchestration still requires a high degree of manual intervention since it is prone to misconfiguration.

Based on the aforementioned aspects, a next generation of container orchestrators needs to integrate a higher level of automation, both for the setup and management of container clusters, as well as during deployment and operation of containerized applications. Ideally, the orchestration of containers should also take into consideration the capability to assist a feasible selection of microservices. For instance, based on application requirements and surrounding context, one could perform data analytics on different locations, eventually selecting different analytics components, such as different classification algorithms, in a way that does not impact application design. Under cluster orchestration, one aspect being addressed concerns supporting dynamic orchestration offloading to address challenges derived from the higher degree of automation in regards to container and workload mobility between nodes in a single cluster and across different clusters, (status synchronization, which information to exchange and disclose, safe handover). In particular, the research is addressing the migration of microservices during runtime to further reduce latency and energy consumption from the perspective of the involved devices. This may be required whenever the behaviour of the application or the infrastructure changes. This

[56] Argerich, Mauricio Fadel, Bin Cheng, and Jonathan Fürst. "Reinforcement learning based orchestration for elastic services." In 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 352–357. IEEE, 2019.

[57] Mao, Hongzi, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. "Resource management with deep reinforcement learning." In Proceedings of the 15th ACM Workshop on Hot Topics in Networks, pp. 50–56. 2016.

[58] D'Eramo, C., Tateo, D., Bonarini, A., Restelli, M., & Peters, J. Sharing Knowledge in Multi-Task Deep Reinforcement Learning. In International Conference on Learning Representations. 2019.

[59] Potvin, P., Nabaee, M., Labeau, F., Nguyen, K. K., & Cheriet, M. Micro service cloud computing pattern for next generation networks. In Smart city 360° (pp. 263–274). Springer, Cham. 2016.

[60] https://www.docker.com/

[61] Bernstein, David. "Containers and cloud: From lxc to docker to kubernetes." IEEE Cloud Computing 1.3: 81–84. 2014.

[62] Mordor Intelligence. Container as a Service Market—Growth, Trends, and Forecast (2020 – 2025). Report ID 4845963, January 2020.

→

may be caused by changes in the load (additional computing tasks, network traffic etc.), changes in network connectivity (caused by mobility of devices), failure of parts of the infrastructure (nodes, network), or even considering the user context and corresponding changes. The advances being considered as extensions to current Kubernetes are threefold. First, our research considers application requirements, infrastructure capabilities and also context information (of the user, or of infrastructure components) when scheduling and/or re-scheduling containerized workloads. Secondly, the aim is to explicitly support scenarios with with a high level of use and device mobility. For this, mobile device platforms need to be supported such that they can be part of the managed container clusters and run containerized workloads. The properties of each part of the infrastructure need to be monitored, updated, and even predicted on a regular basis (connectivity, network conditions, system load). This may be required to support parts of an application running on a mobile device for instance, when the user is going to lose network connectivity, which could be predicted from a given context.

Embedded AI benchmarking

Engineering intelligence in the edge network requires integrating AI distributed methods to better support learning and inference within the decentralized edge network, and relies on machine learning *(ML)* models of which the most promising widely applied in cloud environments require high computing power, energy and memory. This is incompatible with most of the devices that are deployed in far edge infrastructures. Of particular relevancy for next generation IoT applications is the potential for embedded applications to take local advantage of ML models. IoT devices often integrate limited memory, such as 8kB, and their MCUs have a maximum clock frequency of 50MHz, often with no hardware acceleration. Executing a regular ML model is therefore not possible in real-time in IoT environments. One trend of research that focuses on this issue is to rely on embedded AI applications running in smartphones to export the finished model (graph) after training. For instance, a *deep learning* (DL) model is prototyped in a deep learning framework such as Caffe or Tensorflow, but trained on the cloud or a powerful edge controller, often integrating several GPUs. The finished model can then be exported to the far edge personal device. However, it should be noted that today smartphones are powerful computing devices, with storage and computing capability, with battery consumption being the primary constraint. Common IoT devices are even more constrained in terms of storage, memory, computing power and energy dissipation.

Another line of action debated in related literature concerns hardware adaptation, in particular in terms of memory access and usage. In the context of edge computing, such an advantage is still not clear.

fortiss is instead exploring software solutions that can assist in "shrinking ML models". Specifically, fortiss is currently pursuing the benchmarking of AI engineering tools such as TinyML[69] within Industrial IoT infrastructures. The stated goal of TinyML is to bring ML inference to ultra-low-power devices typically under 1 mW[70]. This creates relevant advantages for enabling responsiveness and privacy while overcoming issues with energy consumption in a decentralized edge environment, and is of particular relevance within wireless infrastructures and ultra-low-power devices in industrial environments, which require fast responses (subsecond).

[63] Al Jawarneh, I. M., Bellavista, P., Bosi, F., Foschini, L., Martuscelli, G., Montanari, R., & Palopoli, A. Container Orchestration Engines: A Thorough Functional and Performance Comparison. In ICC 2019-2019 IEEE International Conference on Communications (ICC) (pp. 1–6). IEEE. 2019.

[64] https://kubernetes.io/

[65] https://docs.docker.com/engine/swarm/

[66] http://mesos.apache.org/

[67] https://www.stackrox.com/kubernetes-adoption-and-security-trends-and-market-share-for-containers/

[68] A container cluster consists of one master node, and multiple worker nodes. The master node provides all of the coordination between worker nodes.

[69] https://www.tinyml.org/

[70] Banbury, C. R., Reddi, V. J., Lam, M., Fu, W., Fazel, A., Holleman, J., ... & Yadav, P. (2020). Benchmarking TinyML Systems: Challenges and Direction. arXiv preprint arXiv:2003.04821.

Scientific outcome: publications, software under development, initiatives

Research activities focused on edge AI within industrial IoT environments were initially established in 2020. In this context we highlight the 2020 review paper *"A Review on Scaling Mobile Sensing Platforms for Human Recognition: Challenges and Recommendations for Future Research"* (Carvalho et al, 2020), which addresses challenges for mobile sensing platforms aimed at assisting a future design of these infra-structures, which today are the basis for scenarios such as *mobile crowd sensing*[71]. Of particular relevance to edge AI is raising the awareness of the simplified classification models (due to the limitations of embedded devices) and also the discussion related to classification challenges.

A second outcome concerns the development of an "edge AI ecosystem service" led by fortiss within the context of the "AI on-demand platform for regional interoperable Digital Innovation Hubs Network (H2020 DIH4AI project[72]), where fortiss is in the early stages of developing a living hub in Munich to assist in expanding and empowering edge AI research and innovation in Europe, by offering a set of tools and services focused on industrial IoT aspects. This expansion integrates an open-source edge AI online catalogue that is expected to be available in early 2022 via the DIH4AI portal.

A third outcome concerns the establishment of a demonstrator within the fortiss "Industrial IoT lab"[73] focused on dynamic off-loading of edge AI services within a specific smart cities use case, where a new open-source software (extensions for Kubernetes) under development and scheduled for release in 2021 (*Mobilek8s*) is expected to support the handover of containerized edge AI services based on user context such as roaming.

[71] Guo, Bin, et al. "Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm." ACM computing surveys (CSUR) 48.1 (2015): 1–31.

[72] https://cordis.europa.eu/project/id/101017057.

[73] https://www.fortiss.org/forschung/living-lab/detail/iiot-lab

LITERATURE

Sofia, R. C., Carvalho, L. I., & Pereira, F. M. (2019). The Role of Smart Data in Inference of Human Behavior and Interaction. CRC Press, Big Data Series. Pp 190–211.

Carvalho, L.I., & Sofia, R.C. (2020). A Review on Scaling Mobile Sensing Platforms for Human Activity Recognition: Challenges and Recommendations for Future Research. IoT, 1(2), 451–473.

2.8 Neuromorphic Computing

Authors:
Axel von Arnim, Manos Angelidis

Rationale

Spiking neural networks—the third generation of neural networks

Spiking neural networks (SNN) are considered the third generation of neural networks. First computational models of biological neurons were proposed in the early 50s[74] but simpler models such as the most widely-used Leaky-Integrate-and-Fire neuron found wider adoption in the mid-60s[75]. These kinds of neuron models, which were extensively described[76], were used in proper spiking neural networks in the mid-90s[77]. The term itself appeared in 1996 in a now famous publication[78]. Unlike artificial (conventional) neural networks, SNNs mimic the way that biological neurons work, more or less accurately from the biological standpoint, depending on the application[79], by exchanging data in "spike" format (impulses) and simulating biological neuron parameters such as time constants or membrane voltage. This primarily non-linear detailed behavior and the necessary conversion of data to spikes adds a layer of complexity that at first glance is an unnecessary cost. Crucial advantages have nevertheless surfaced as research advanced. First, spike communication is sparse by nature[80], allowing for less computing activity in spiking neural networks than in conventional ones. Second, contrary to artificial neural networks, SNNs integrate time at the neuron level by construction, since spiking neurons are modelled from biological neurons, which have natural time constants. This is an enormous advantage when processing time dependent data, such as sensor data[81], speech[82], audio[83] or video[84]. Third, dedicated neuromorphic hardware is being released constantly by research and industry, such as SpiNNaker and Intel's neuromorphic chip codenamed Loihi[85], which dramatically accelerates SNN, allowing them to compete with conventional networks in terms of computing speed[86]. Fourth, non linearities in computation, which first prevented the use of backward propagation for training purposes, are being overcome as more and more alternative learning algorithms are produced by research[87 88]. Fifth, and best, energy consumption in SNN is dramatically reduced compared to conventional networks, by orders of magnitude, due to the dedicated efficient neuromorphic hardware andthe natural sparsity of SNNs.

At fortiss, we are accelerating the transfer of artificial network-based use cases to SNNs so that industry can take advantage of these five assets in the future. This is what we describe in the paragraph below.

[74] A.L. Hodgkin, A.F. Huxley (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol

[75] Geisler CD, Goldberg JM (1966) A stochastic model of the repetitive activity of neurons. Biophys J 6:53–69

[76] Gerstner, W., & Kistler, W. (2002). Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge: Cambridge University Press

[77] Gerstner W (1995) Time structure of the activity in neural network models. Phys Rev E 51:738–758

[78] Wolfgang Maas. (1997). Networks of spiking neurons: the third generation of neural network models. Trans. Soc. Comput. Simul. Int. 14, 4 (Dec. 1997), 1659–1671

[79] Tavanaei A, Ghodrati M, Kheradpisheh SR, Masquelier T, Maida A (2018). Deep learning in spiking neural networks. Neural Netw. 2019 Mar;111:47–63. Epub 2018 Dec 18

[80] Tang, P.T.P., Lin, T., Davies, M. (2017): Sparse coding by spiking neural networks: convergence theory and computational results. CoRR abs/1705.05475

[81] Stromatias E, Soto M, Serrano-Gotarredona T, Linares-Barranco B (2017). An Event-Driven Classifier for Spiking Neural Networks Fed with Synthetic or Dynamic Vision Sensor Data. Frontiers in Neuroscience

→

Neuromorphic Computing—AI for edge and mobile computing

The first neuromorphic computing chips emerged as an attempt to better understand how basic principles of biological neuronal networks give rise to perception and motor control[89]. Only recently up-scaling of hardware neuromorphic chips such as IBM's TrueNorth[90], Brainchip's Akida[91] or academic boards SpiNNaker[92] or BrainScale allowed researchers to explore them as a competitive tool for solving AI tasks. Intel's neuromorphic research chip Loihi is today's most advanced neuromorphic chip, which allows implementation of large-scale spiking neuronal networks with state-of-the-art results in massively parallel search[93], fast sensory learning[94,95], and adaptive motor control[96,97]. This hardware, with various architectures and design (from fully digital to fully analog modelling of spiking neurons) allows SNNs to be executed and increasingly trained on chip. This paves the way for the long-awaited adoption of modern AI in mobile and edge computing.

In the field of mobile robotics, neuromorphic hardware and SNNs are expected to show dramatic improvements in adaptive locomotion control (in particular online learning, while the device is operating) with regards to energy efficiency, which is key. fortiss is investigating dedicated research line for this topic and is showing promising results in motion control.

Smart sensors can also take huge advantage of neuromorphic hardware when they can speak the same language : spikes. This is the case for more and more sensors that deliver native spiking data, in particular so-called event-based cameras, also known as dynamic vision sensors (DVS). These cameras grab visual information in the form of a continuous flow of pixel intensity events instead of the traditional frames. This makes an enormous difference in terms of latency, as such sensors deliver data in real-time, not in periodic frames. Of course, their natural data sparsity and the energy efficiency of neuromorphic hardware to process this data again serves to disrupt the energy efficiency issues in conventional vision sensors. fortiss is considering this disruptive trend in real-time image processing as a dedicated research line.

Scientific approach

1) Embodied AI and efficient motion control in neuromorphic computing

State-of-the-art

Robust, adaptive and intelligent autonomous machines are one of the major breakthroughs promised by the ongoing AI revolution. Autonomous agents acting with minimal supervision, enhanced locomotive skills and low-energy consumption can pave the way for multiple industry applications, such as medical assistants, warehouse autonomous vehicles, delivery drones and space exploration among others. To accomplish such tasks, autonomous locomotion is a crucial factor[98]. Our partners, Prof. Auke Ijspeert's Lab in EPFL[99] and the research teams at the Human Brain Project[100] have been pioneers in shedding light on this question and in developing technologies to address it. Despite the impressive progress achieved over the last decades in terms of autonomous machines, some key challenges remain unaddressed. How can we build machines that can present adaptive capabilities (as in animals) when it comes to motion in varying conditions and fine-movement control? Existing models have tried to address such questions, often providing solutions limited in scope[101], specialized[102] or

[82] Wu, J., Yılmaz, E., Zhang, M., Li, H., & Tan, K. C. (2020). Deep Spiking Neural Networks for Large Vocabulary Automatic Speech Recognition. Frontiers in neuroscience, 14, 199

[83] Pan, Z., Chua, Y., Wu, J., Zhang, M., Li, H., & Ambikairajah, E. (2020). An Efficient and Perceptually Motivated Auditory Neural Encoding and Decoding Algorithm for Spiking Neural Networks. Frontiers in neuroscience, 13, 1420

[84] Panda, P., & Srinivasa, N. (2018). Learning to Recognize Actions From Limited Training Examples Using a Recurrent Spiking Neural Model. Frontiers in Neuroscience, 12, 1

[85] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Prasad Joshi, Andrew Lines, Andreas Wild, Hong Wang (2018). Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. IEEE Micro PP(99):1-1

[86] Peter Blouw, Xuan Choo, Eric Hunsberger, & Chris Eliasmith. (2019). Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware

[87] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass (2018). Long short-term memory and learning-to-learn in networks of spiking neurons. 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montreal, Canada

[88] E. O. Neftci, H. Mostafa, F. Zenke (2019). Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks . IEEE Signal Processing Magazine 36(6):51-63

→

difficult to tune and train[103]. On the contrary biological organisms seem to solve these problems effortlessly all with the same basic infrastructure and building blocks, which happen to be … natural spiking neural networks.

Central pattern generators (CPG) are specialized neural circuits that with minimal supervision from other areas of the nervous system can generate synchronized and organized electrical signals to the muscles controlling both involuntary motions (breathing, swallowing) as well as voluntary motions. These properties are all highly desired from the perspective of autonomous locomotion. Many systems have been developed that make use of CPG models as their controllers, most notably the Salamandra Robotica from EPFL[104]. Still, a spiking CPG that could be run on energy-efficient embeddable neuromorphic hardware has been lacking to date.

Research

Our contribution to the scientific challenge of adaptive locomotion was made possible by a very fruitful collaboration with EPFL and Intel's Neuromorphic Research Community. First, we came up with a novel spiking CPG model that can run on multiple neuromorphic platforms. This model, in comparison to other spiking CPG models that are based on low-level biological information and which are system-specific, offers unique properties that make it a good candidate for autonomous locomotion under neuromorphic control. Indeed, the output of the coupled oscillators of the CPG are synchronized, robust to external perturbations and easy to control with high level drives. We implemented our algorithm on a virtual lamprey-like model in a simulated experiment in the Neurorobotics Platform (Falotico et al, 2017) (NRP).

The NRP is a simulation software combining virtual embodiment and neuromorphic computing that fortiss develops within the framework of the Human Brain Project. It proves to be extremely helpful for designing, testing and benchmarking our SNN based algorithms on various neuromorphic hardware. It supports spiking simulators such as the Neural Engineering Framework[105] and its software component Nengo and interfaces with Intel's neuromorphic research chip Loihi and the SpiNNaker neuromorphic board. When it comes to models like the lamprey specifically, which evolves in water, most physical simulation platforms do not offer realistic fluid simulation. To address this problem we have complemented the NRP with fluid dynamics based on the smoothed particle hydrodynamics method[106]. This method makes use of discretized particles that carry physical properties such as mass and energy as they move in 3D space, and can be used to solve the 3D Navier-Stokes equations that describe fluid flow. This enables the interaction of water physics with virtual embodiments, leading in our case to a very realistic simulated locomotion of our lamprey model in water, controlled with our spiking CPG.

We used our model to investigate the performance of neuromorphic hardware in real-time and to showcase their capabilities in terms of energy efficiency and computational speed. We managed to show the impressive performance of neuromorphic hardware compared to the CPU when running spiking neural networks, with energy performance three orders of magnitude better than CPUs and with computational speed advantages (Figure 14). This result is a demon-stration of the real-time performance of neuromorphic hardware which proves their usefulness in locomotion control. Our second contribution is the simulation of the

[89] Farquhar, Ethan; Hasler, Paul. (2006). A field programmable neural array. IEEE International Symposium on Circuits and Systems. pp. 4114–4117

[90] Modha, Dharmendra (2014). "A million spiking-neuron integrated circuit with a scalable communication network and interface". Science. 345 (6197): 668–673

[91] https://brainchipinc.com

[92] Xin Jin; Furber, Steve B.; Woods, John V. (2008). "Efficient modelling of spiking neural networks on a scalable chip multiprocessor". 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 2812–2819

[93] Paxon Frady, E., Garrick Orchard, David Florey, Nabil Imam, Ruokun Liu, Joyesh Mishra, Jonathan Tse, et al.. Sommer, and Mike Davies (2020). "Neuromorphic Nearest-Neighbor Search Using Intel's Pohoiki Springs." arXiv

[94] Taunyazov, T., Sng, W., See, H. H., Lim, B., Kuan, J., Ansari, A. F., … & Soh, H. (2020). Event-driven visual-tactile sensing and learning for robots. arXiv preprint arXiv:2009.07083

[95] Imam, N., & Cleland, T. A. (2020). Rapid online learning and robust recall in a neuromorphic olfactory circuit. Nature Machine Intelligence, 2(3), 181–191

[96] DeWolf, T., Jaworski, P., & Eliasmith, C. (2020). Nengo and low-power AI hardware for robust, embedded neurorobotics. Frontiers in Neurorobotics, 14

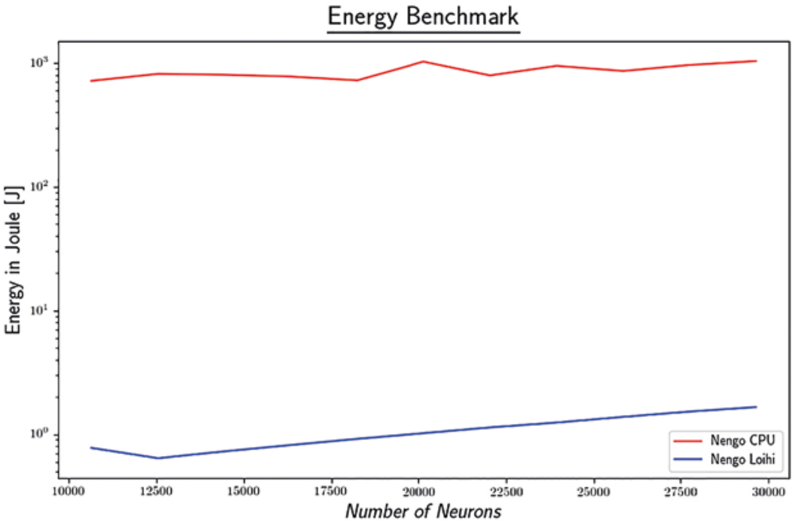


Figure 14.
Energy consumption comparison between execution on cpu and loihi

spiking CPG model into the NRP, and the testing of hypotheses on locomotion control with neuromorphic hardware. We show how our model can efficiently control a simulated lamprey in speed and direction, and with the appropriate feedback from the environment, also overcome obstacles. Furthermore, we present a unified fluid dynamics simulation with a controller based on our CPG model that can generate self-propelled locomotion, by applying the fluid forces to the model. These achievements are showcased in a video (Angelidis and von Arnim, 2020) and our a written document (Angelidis et al, 2021).

In another project going on at the time of this writing, we are investigating adaptive locomotion control on neuromorphic hardware in a different use case, yet sharing the same energy latency constraints: industrial arm control for object insertion. We expect great progress in motion learning from real-time sensory feedback by taking advantage of the adequation between spiking neural networks and timely sensor events.

[97] Stagsted, R. K., Vitale, A., Binz, J., Larsen, L. B., & Sandamirskaya, Y. (2020). Towards neuromorphic control: A spiking neural network based PID controller for UAV. ROBOTICS: SCIENCE AND SYSTEMS XVI

[98] Rubio, F., Valero, F., & Llopis-Albert, C. (2019). A review of mobile robots: Concepts, methods, theoretical framework, and applications. International Journal of Advanced Robotic Systems

[99] https://www.epfl.ch/labs/biorob/

[100] https://www.humanbrainproject.eu/en/science/overview/

[101] M. A. Lewis, F. Tenore and R. Etienne-Cummings (2005), "CPG Design using Inhibitory Networks," Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp. 3682–3687

[102] K. Inoue, Shugen Ma, and Chenghua Jin (2004). Neural oscillator network-based controller for meandering locomotion of snake-like robots. In IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, volume 5, pp. 5064–5069

[103] Zhelong Wang, Qin Gao, and Hongyu Zhao (2017). CPG-Inspired Locomotion Control for a Snake Robot Basing on Nonlinear Oscillators. Journal of Intelligent & Robotic Systems, 85(2):209–227

→

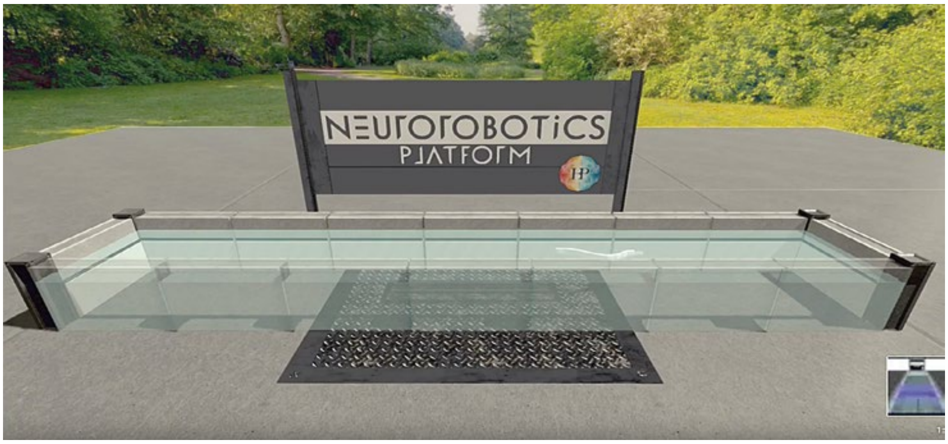


Figure 15.
The simulated lamprey model in the neurorobotics platform

2) Event-based efficient perception in neuromorphic computing

State-of-the-art

In mobile applications such as advanced driving assistance systems, there is a clear need to solve the challenge of the growing presence of AI services that increase the required computing power, their complexity and energy consumption. Another critical issue for vision-based sensing is the uncompressible latency of frame-based visual sensors. A full frame must be grabbed before any calculation is started. Differential information, like the speed of detected moving objects, is even more delayed because it needs a series of grabbed frames, making important security decisions to be taken in the past. Industry thus needs a long-term solution for the efficient integration of artificial intelligence in applications with low-latency sensing.

We propose the use of neuromorphic hardware and event-based cameras (EBC) to implement low-energy embedded object detection and tracking sensors with embedded AI processing and very low latency optical flow calculation. The advantage of EBCs over conventional frame-based cameras is that they deliver pixel intensity changes on the fly (events). They speak the same data language (spikes) as spiking neural networks, which run on accelerated neuromorphic hardware and implement sparse and energy efficient object detection and tracking algorithms.

Very few works deal with the end-to-end consistent vision[107] where events and time are considered from the off-line learning phase to the on-line embedded prediction. The use of spiking neural networks in the state-of-the-art technology can thus be summarized in three approaches: transposition from formal networks that have learned through supervised learning[108], adaptation of back-propagation algorithms to the case of SNN by modifying the transfer functions of each neuron to make them derivable (time-coding[109] or surrogate gradient[110]), and finally by using an unsupervised learning rule[111]. But in the vast majority of cases, these approaches only apply to frame-based data or shallow networks.

Optical flow determines the motion of objects while taking into account the relative motion between an observer and the scene. It can be estimated accurately by solving partial differential equations with the iterative Horn-Schunk method[112]. Asynchronous event-based optical flow has been employed in neu-

[104] A. J. Ijspeert, A. Crespi, D. Ryczko, and J.-M. Cabelguen (2017). From Swimming to Walking with a Salamander Robot Driven by a Spinal Cord Model. Science, 315(5817):1416–1420

[105] Chris Eliasmith and Charles Anderson (2004). Neural engineering: Computation, representation, and dynamics in neurobiological systems. IEEE Transactions on Neural Networks, 15(2):528–529, March 2004

[106] Liu, M.B., Liu, G.R (2010). Smoothed Particle Hydrodynamics (SPH): an Overview and Recent Developments. Arch Computat Methods Eng 17, 25–76

[107] J. Li, S. Dong, Z. Yu, Y. Tian and T. Huang (2019), Event-Based Vision Enhanced: A Joint Detection Framework in Autonomous Driving, IEEE International Conference on Multimedia and Expo

[108] N. Abderrahmane, E. Lemaire and B. Miramond (2019). Design Space Exploration of hardware spiking neurons for embedded Artificial Intelligence. Elsevier Journal on Neural Networks

→

romorphic hardware (IBM TrueNorth[113], Loihi[114]) as well as with EBC[115]. Optical flow computation has been evaluated with industrial EBC, including a translating cart. Pre-processing methods such as smoothing led to improved accuracy. Although many authors focus on optical flow computation in SNNs, few use it in vision with neuromorphic hardware.

Research

fortiss will explore the estimation of optical flow from the approach of groups of diffeomorphisms, which are part of variational methods[116]. In particular, they give hope for spiking compatibility.

→ Research question 1:
How to model optical flow solutions in a Lie group? What properties can be derived? Concerning learning methods, the closeness of the computed functions with the diffeomorphisms obtained by variational methods is to be determined. The smoothness and the generalization power of the model are to be evaluated. Another investigation is the use of a temperature parameter, in contrast to contrastive learning, in order to relax the similarity measure.

→ Research question 2:
To what extent does the computed optical flow coming from learning methods compare to variational methods (accuracy, convergence, latency)? Optical flow eases motion detection and action recognition. They are typically performed on devices with limited resources such as drones or mobile devices. For privacy and efficiency reasons, neuromorphic hardware and sensors are embedded. The scalar to binary conversion and the integration into the spiking neural network framework are to be thoroughly evaluated.

→ Research question 3:
How to integrate optical flow in an event-based detection architecture?

These three questions are at the heart of our research line and will have applications in advanced driving assistance systems (road object detection and tracking) as well as in industrial automation (industrial arm adaptive control with visual feedback).

[109] S. B. Shrestha and G. Orchard (2018). {SLAYER}: Spike Layer Error Reassignment in Time. NIPS

[110] S. R. Kheradpisheh and T. Masquelier (2020). Temporal backpropagation for spiking neural networks with one spike per neuron. International Journal of Neural Systems

[111] M. Mozafari, M. Ganjtabesh, A. Nowzari-Dalini, S. Thorpe and T. Masquelier (2019). Bio-inspired digit recognition using reward modulated spike-timing-dependent plasticity in deep convolutional networks. Pattern Recognition 94

[112] B. Horn and B. Schunck (1981), Determining Optical Flow, Techniques and Applications of Image Understanding

[113] G. Haessig, A. Cassidy, R. Alvarez, R. Benosman and G. Orchard (2017), Spiking Optical Flow for Event-based Sensors Using IBM's TrueNorth Neurosynaptic System, IEEE Transactions on Biomedical Circuits and Systems

[114] F. Paredes-Vallés, K. Scheper and G. de Croon (2019), Unsupervised Learning of a Hierarchical Spiking Neural Network for Optical Flow Estimation: From Events to Global Motion Perception, IEEE Transactions on Pattern Analysis and Machine Intelligence

[115] M. Liu and T. Delbruck (2017), Block-Matching Optical Flow for Dynamic Vision Sensors: Algorithm and FPGA Implementation, IEEE International Synopsium on Circuits and Systems

[116] G. Sundaramoorthi and A. Yezzi (2018). Accelerated Optimization in the PDE Framework: Formulations for the Manifold of Diffeomorphisms. Arxiv preprint

→

Research results

fortiss has delivered promising results in spiking adaptive locomotion control:

- Demonstrators: A simulated spiking locomotion demonstrator running in the Neurorobotics Platform
- Software: A version 3.0 of the Neurorobotics Platform (NRP) (refer to Neurorobotics) that provides particle-based fluid simulation for realistic environment feedback on moving agents
- Publications:
  - On adaptive locomotion control (Vandesompele et al, 2019; Angelidis et al, 2021; Allegra Mascaro et al, 2020)
  - On simulated embodiment for neuromorphic computing research (Falotico et al, 2017; Vannucci et al, 2015; Matthes et al, 2019; Bornet et al, 2019; Capolei et al, 2019; Tieck et al, 2019).
  - Videos on adaptive locomotion control(Angelidis and von Arnim , 2020), NRP release 3.0 (von Arnim et al, 2020), NRP project management (NRP fortiss, 2020)
- Research partnerships with large industrial players in neuromorphic hardware and event-based vision

As part of the Human brain Project and the Intel Neuromorphic Research Community, fortiss is strengthening its neuromorphic computing research lines and paving the way for the adoption of spiking neural networks in real world use cases, through fundamental research, applied use cases, testing and the benchmarking of neuromorphic computing techniques against the state-of-the-art in AI.

LITERATURE

Angelidis, M., von Arnim, A. (2020): https://www.youtube.com/watch?v=ThhhixVDy4w.

Angelidis, E., Buchholz, E., O'Neil, J. P. A., Rougè, A., et al. (2021). A Spiking Central Pattern Generator for the control of a simulated lamprey robot running on SpiNNaker and Loihi neuromorphic boards. arXiv preprint arXiv:2101.07001.

Allegra Mascaro, A. L., Falotico, E., Petkoski, S., et al (2020). Experimental and Computational Study on Motor Control and Recovery After Stroke: Toward a Constructive Loop Between Experimental and Virtual Embodied Neuroscience. Frontiers in systems neuroscience, 14, 31.

Bornet, A., Kaiser, J., Kroner, A., Falotico, E., Ambrosano, A., Cantero, K., ... & Francis, G. (2019). Running Large-Scale Simulations on the Neurorobotics Platform to Understand Vision–The Case of Visual Crowding. Frontiers in neurorobotics, 13, 33.

Capolei, M. C., Angelidis, E., Falotico, E., Lund, H. H., & Tolu, S. (2019). A biomimetic control method increases the adaptability of a humanoid robot acting in a dynamic environment. Frontiers in neurorobotics, 13, 70.

Falotico, E., Vannucci, L., Ambrosano, A., Albanese, U., Ulbrich, S., Vasquez Tieck, J. C., ... & Gewaltig, M. O. (2017). Connecting artificial brains to robots in a comprehensive simulation framework: the neurorobotics platform. Frontiers in neurorobotics, 11, 2.

Matthes, C., Weissker, T., Angelidis, E., Kulik, A., Beck, S., Kunert, A., ... & Froehlich, B. (2019). The Collaborative Virtual Reality Neurorobotics Lab. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 1671–1674). IEEE.

Neurorobotics. http://neurorobotics.net.

NRP fortiss (2020). https://www.youtube.com/watch?v=jDf-vejNQi4.

Tieck, J. C. V., Kaiser, J., Steffen, L., et al. (2019). The Neurorobotics Platform for Teaching–Embodiment Experiments with Spiking Neural Networks and Virtual Robots. In 2019 IEEE International Conference on Cyborg and Bionic Systems (CBS) (pp. 291–298). IEEE.

Vandesompele, A., Urbain, G., Mahmud, H., & Dambre, J. (2019). Body randomization reduces the sim-to-real gap for compliant quadruped locomotion. Frontiers in neurorobotics, 13, 9.

Vannucci, L., Ambrosano, A., Cauli, N., et al. (2015). A visual tracking model implemented on the iCub robot as a use case for a novel neurorobotic toolkit integrating brain and physics simulation. In 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids) (pp. 1179–1184). IEEE.

Von Arnim, A. et al. (2020). https://www.youtube.com/watch?v=VF3xd3mXTqY&list=PLFfa5EHopIFoz5xFlnF5SGr2mdUYqb93Z.

> Despite technological advances that have led to the proliferation of AI-based solutions, questions remain about the level of trust that can be placed in AI systems. What is missing, therefore, is a rigorous approach to building and operating AI systems in which people can trust.

3

APPLICATIONS

APPLICATIONS

3.1 fortiss Labs

Author:
Dr. Markus Duchon

Using industry-specific demonstrators as a basis, research and innovation projects for prototyping, testing and industrial use are carried out in the fortiss labs. In accordance with our motto: "Researching, Applying, Shaping", we illustrate current research results and their practical application possibilities in different domains. In this way, we are able to show interested visitors, researchers, application partners and networks how we can shape future developments and exploit the potential associated with digitization. With our labs we make research tangible and experiential. The main vision and mission includes:

- providing suitable domain specific environments for exploration, demonstration and training
- amplifying interdisciplinary work with internal and external organizations as well as with strategic partners, academia and industry
- developing, validating and demonstrating core scientific mechanisms and research results using physical demonstrators and industrial use cases
- addressing and solving real-world problems and gaining visibility as an application oriented research institute—"we do things—and not just talk about them"
- providing an appealing dissemination of research results to a broad audience
- offering interesting platforms for lab courses, student theses and adapt assets to address research and industry challenges in corresponding projects

fortiss currently operates the following domain-specific labs: Industrial IoT, Robotics, Energy, Mobility and Drone. Each of the labs is outlined in more detail over the following pages.

3.1.1 Industrial IoT Lab

Author:

Prof. Dr. Rute Sofia

The IIoT lab covers an end-to-end perspective, from the field-level to the cloud, as represented in Figure 16. The main demonstrators correspond to those that already exist, while the additional ones represent those currently under development.

The IIoT lab holds a set of individual, mobile demonstrators based on open-source software being developed in the IIoT competence center and which, when interconnected, provide an end-to-end perspective of edge-cloud mechanisms that are useful in the context of Industrial IoT. The lab is not intended as a large-scale platform. Instead, it is envisioned to be a forefront, neutral and open lab, with planned interconnectivity to existing IoT large-scale platforms, such as EdgeNet[117], FIT-IoT[118], Named-data Networking[119].

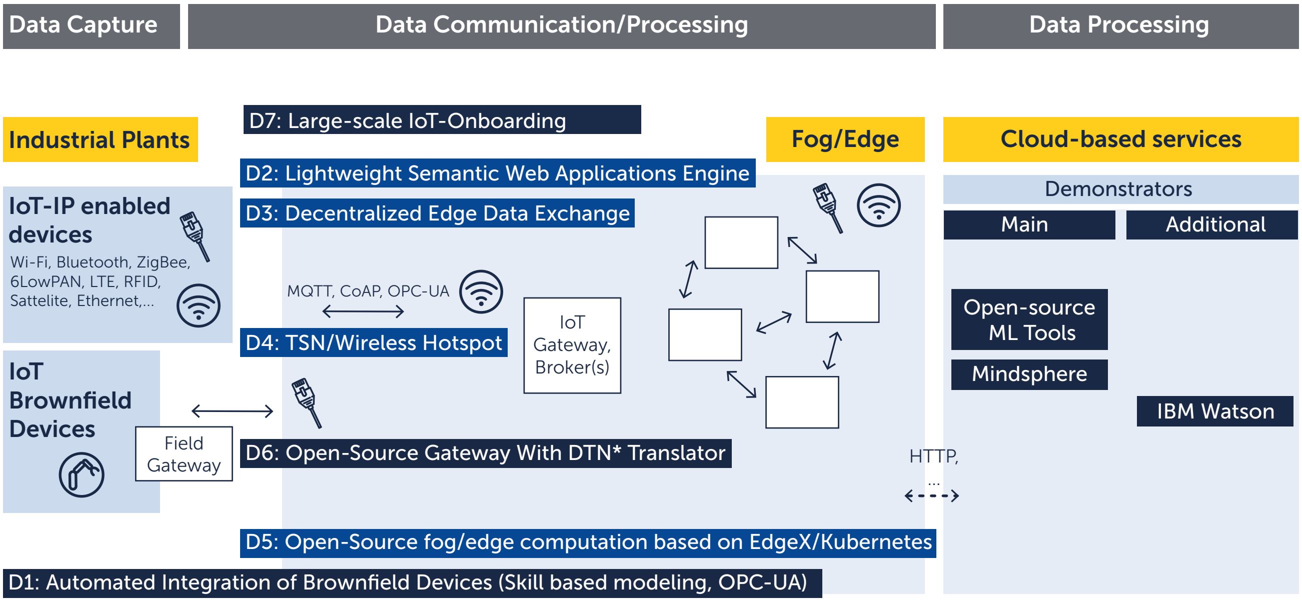

Data Capture
Data Communication/Processing
Data Processing
Industrial Plants
D7: Large-scale IoT-Onboarding
Fog/Edge
Cloud-based services
D2: Lightweight Semantic Web Applications Engine
D3: Decentralized Edge Data Exchange
IoT-IP enabled devices
Wi-Fi, Bluetooth, ZigBee, 6LowPAN, LTE, RFID, Sattelite, Ethernet,...
Demonstrators
Main
Additional
MQTT, CoAP, OPC-UA
IoT Gateway, Broker(s)
D4: TSN/Wireless Hotspot
Open-source ML Tools
Mindsphere
IBM Watson
IoT Brownfield Devices
Field Gateway
D6: Open-Source Gateway With DTN* Translator
HTTP, ...
D5: Open-Source fog/edge computation based on EdgeX/Kubernetes
D1: Automated Integration of Brownfield Devices (Skill based modeling, OPC-UA)

Figure 16.
High-level perspective of the IIoT lab.

In 2020, the lab contemplated the development of 4 different demonstrators and specific open-source development, as illustrated in Figure 17. BFThing (Dorofeev et al, 2020) is open-source middleware being developed by fortiss which provides a way for a legacy device to be integrated into open-source IIoT systems via an automated PLC description into a Web of Things Description (WoT TD) format. Via this novel software module, an edge/fog device, or an IIoT gateway is expected to support bi-directional connectivity to brownfield devices. Standardized communication protocols and data models from IIoT domains as well as conversion tools to integrate legacy devices facilitate the connection. Thereby, smooth and seamless connectivity is established.

[117] https://edge-net.org/

[118] https://www.iot-lab.info/

[119] https://named-data.net/ndn-testbed/

Figure 17.
IIoT lab 2020 demonstrators.

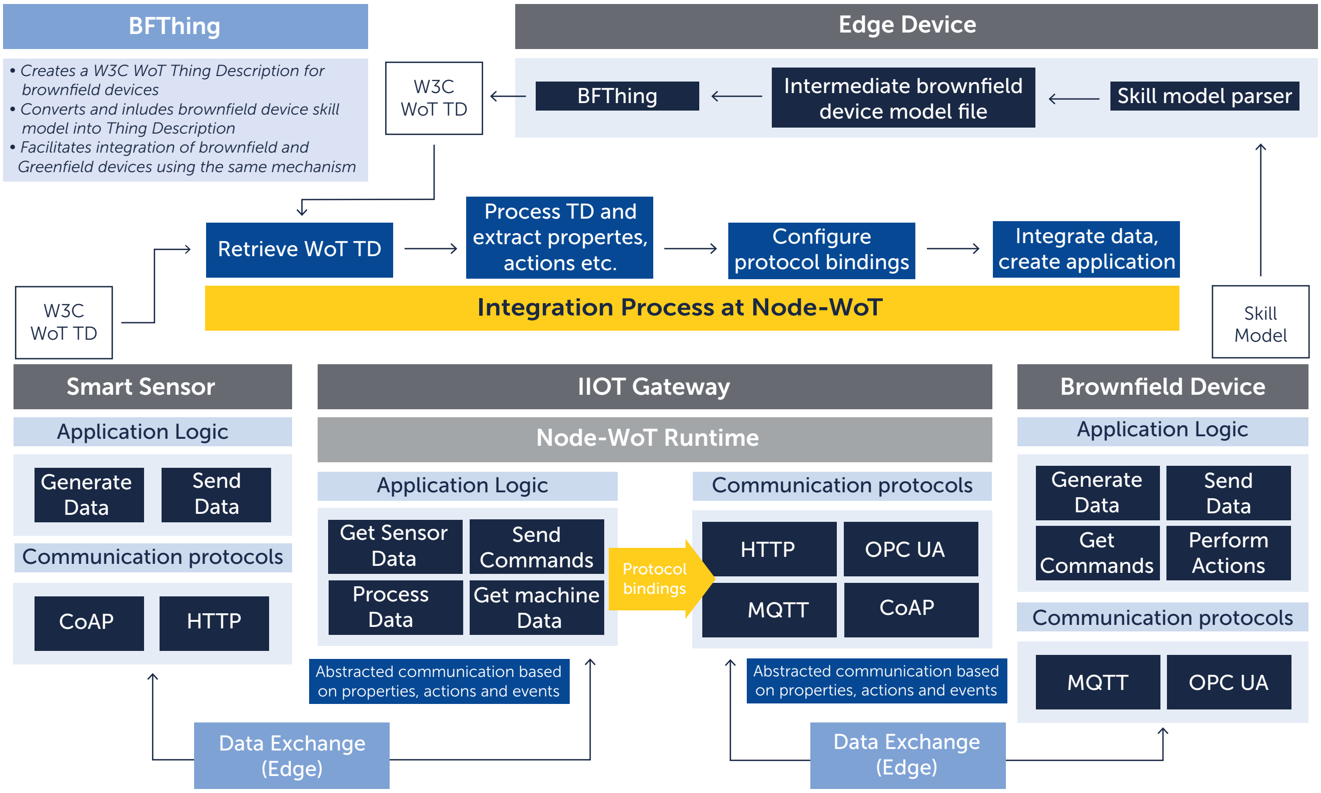
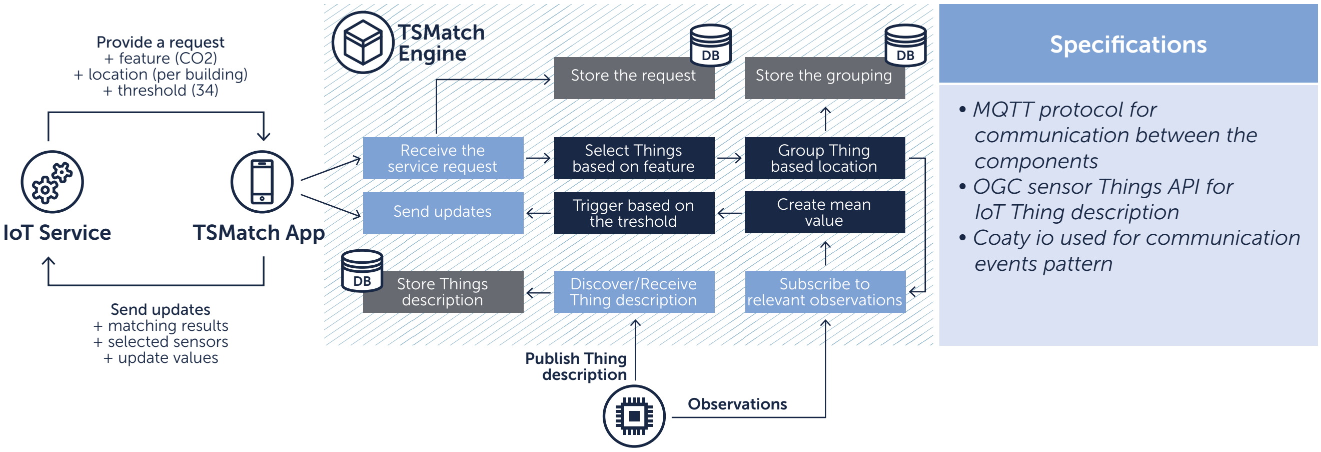

BFThing
• Creates a W3C WoT Thing Description for brownfield devices
• Converts and inludes brownfield device skill model into Thing Description
• Facilitates integration of brownfield and Greenfield devices using the same mechanism
Edge Device
W3C WoT TD
BFThing
Intermediate brownfield device model file
Skill model parser
Retrieve WoT TD
Process TD and extract propertes, actions etc.
Configure protocol bindings
Integrate data, create application
W3C WoT TD
Integration Process at Node-WoT
Skill Model
Smart Sensor
Application Logic
Generate Data
Send Data
Communication protocols
CoAP
HTTP
IIOT Gateway
Node-WoT Runtime
Application Logic
Get Sensor Data
Send Commands
Process Data
Get machine Data
Protocol bindings
Communication protocols
HTTP
OPC UA
MQTT
CoAP
Abstracted communication based on properties, actions and events
Abstracted communication based on properties, actions and events
Data Exchange (Edge)
Data Exchange (Edge)
Brownfield Device
Application Logic
Generate Data
Send Data
Get Commands
Perform Actions
Communication protocols
MQTT
OPC UA

Figure 18.
IIoT BFThing demonstrator.

The TSMatch open-source software supports an automated match between a service (functional and non-functional requirements) and an existing IoT infrastructure, by selecting an optimal set of IoT data sources that can fullfil the desired service requirements. This is performed via 2 components: 1) an IoT App to be installed in an end-user devices, which provides an interface to describe the requi-



- *MQTT protocol for communication between the components*
- *OGC sensor Things API for IoT Thing description*
- *Coaty io used for communication events pattern*

Figure 19.
TSMatch IIoT high-level representation.

rements and view the matching results and updates from an existing IoT infrastructure (sets of sensors and actuators, coined in related literature as "Things"); 2) a server-side component which supports the automated and dynamic matching between the available Things and the service requirements' description provided by the end-user. The TSMatch server component can be run, for instance, on an IoT gateway, end-user device, edge controller or the cloud.

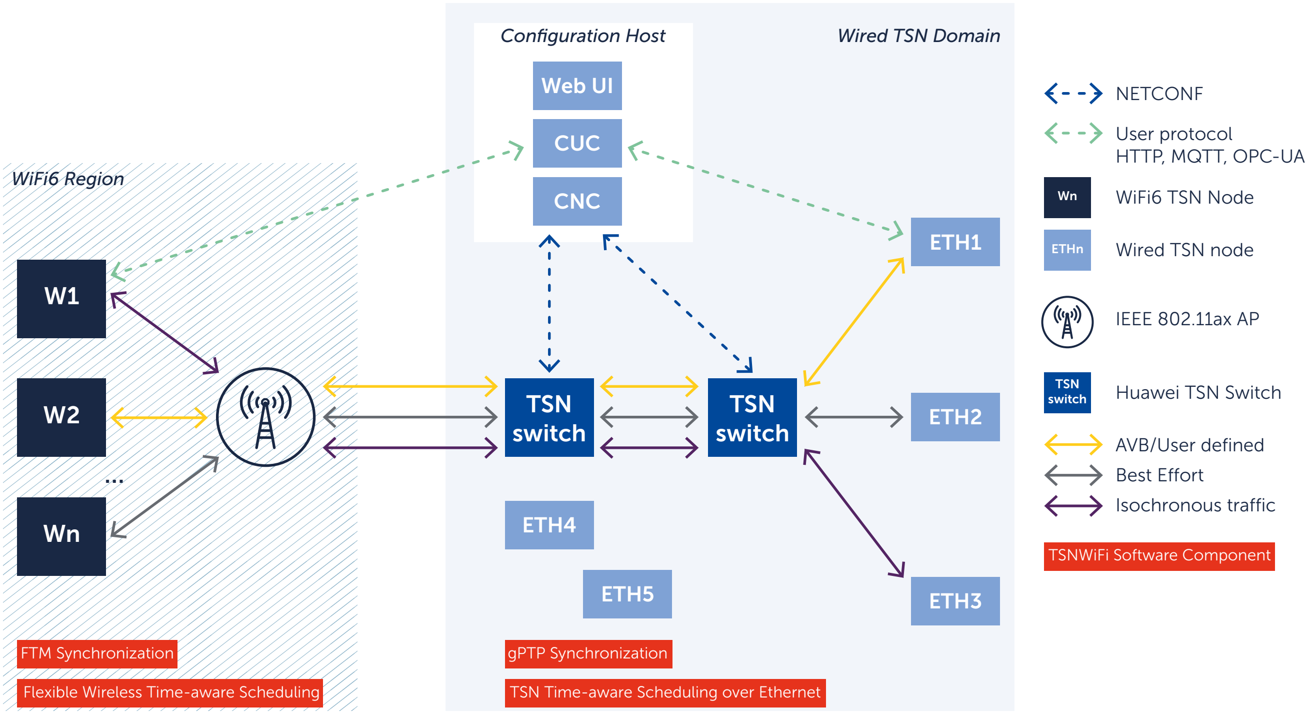


Figure 20.
IIoT TSNWiFi demonstrator.

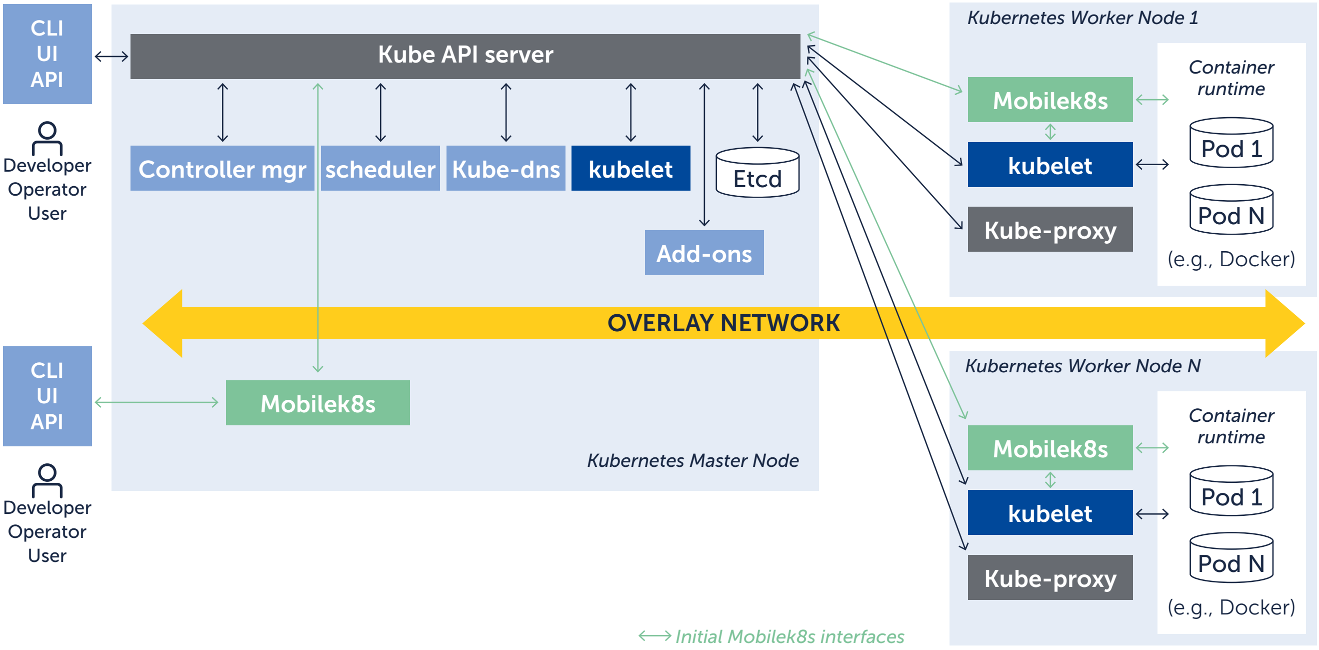
The TSNWiFi demonstrator (see Figure 18, 19, 20, 21) has been set to enhance wireless communications (IEEE 802.11ax) with deterministic capabilities. The demonstrator provides a hybrid wireless/wired TSN infrastructure with multip-

Figure 21. TSNwifi demonstrator under development, partial equipment (bbb, huawei switches, wifi 6 xiaomi AR3600 AP, STAS).

le endpoints (TSN talkers and listeners). Specific software under development concerns extensions based on fine time management (FTM), defined in IEEE 802.1mc for fine-grained localization purposes and currently being integrated in p802.1AS-rev as a mechanism capable of supporting fine-grained synchronization and respective mapping to gPTP on the wired network and flexible wireless time-aware scheduling mechanisms to assist in delivery of end-to-end TSN traffic profiles with guaranteed low latency, zero packet loss, and low latency.

Movek8s corresponds to an extension of Kubernetes that, via context-awareness, behavior learning and inference, improves the orchestration of containerized applications across different edge infrastructures, and is thus expected to reduce the need for human intervention, and to provide support for edge node and service mobility.

Figure 22. Mobilek8s demonstrator under development.



LITERATURE

K. Dorofeev, H. Walzel, R. C. Sofia. HYPERLINK https://www.fortiss.org/fileadmin/user_upload/Veroeffentlichungen/Informationsmaterialien/fortiss_whitepaper_Brownfield_devices_in_IIoT_web.pdf Brownfield devices in IIoT, automatic the integration via semantic technologies. fortiss GmbH Technical White Paper, 2020. ISSN print 2699-1217.

3.1.2 Robotics Lab

Authors:
Alexander Perzylo, Dr. Markus Rickert

The overall goal of the fortiss Robotics Lab is to provide a foundation for research and innovation related to robot-based automation solutions and collaboration with interested stakeholders. It aims to address real-world problems and transfer the latest academic achievements into industrially relevant demonstration platforms and use cases. The implemented showcases are used to evaluate, validate and disseminate research results to a broad audience ranging from industrial partners and other academic institutions to interested students.

They also serve as an open platform for discussions and joint developments on applied research. The focus is on robotic systems engineering, in particular the semantic interoperability of manufacturing resources for knowledge-based autonomous production, as well as model-based software development for robotics and the integration of research results into a common platform.

There are three core research topics. Robot program synthesis based on declarative goal definitions is aimed at helping application domain experts specify production goals at a higher abstraction level and in a familiar language. Using formal knowledge modeling techniques, knowledge from the automation and application domains is semantically described in a machine-interpretable format. Gaps and ambiguities in potentially underspecified instructions are closed via logical inference and planning. As a result, a fully parameterized robot program can be automatically synthesized.

Semantic interoperability in cyber-physical production systems (CPPS) deals with research into approaches for the flexible reconfiguration of heterogeneous CPPS based on task specifications. This includes the automatic reconfiguration of software and hardware components via semantic resource models, as well as the matching of semantically modeled capabilities with formal requirements derived from the manufacturing processes and the associated products.

Model-based software engineering for robotics develops seamless systems engineering approaches, from low-level control to high-level planning. The research covers the abstraction of heterogeneous software and hardware components as a foundation for a hardware-agnostic system architecture and an open-source implementation within the robotics library. Real-word demonstrators are essential for evaluating the applicability of these research topics in relevant use cases and for their dissemination to the public. The Robotics Lab currently features the following demonstators.

Automated configuration of robots & analytics (Figure 23) combines semantic process knowledge and machine learning based data analytics techniques to increase the accessibility of self-monitoring robot systems for manufacturing SMEs. As technical systems are always subject to occasional errors, complex robot systems in particular must be able to cope with anomalies and potential failures during production. A barrier for machine learning based approaches to anomaly detection in production is the need for large amounts of data that is often manually trained. Our goal is to automatically lable process data based on semantic knowledge from manufacturing resources and the automation and application domains. Suitable anomaly indicators can then be derived from the

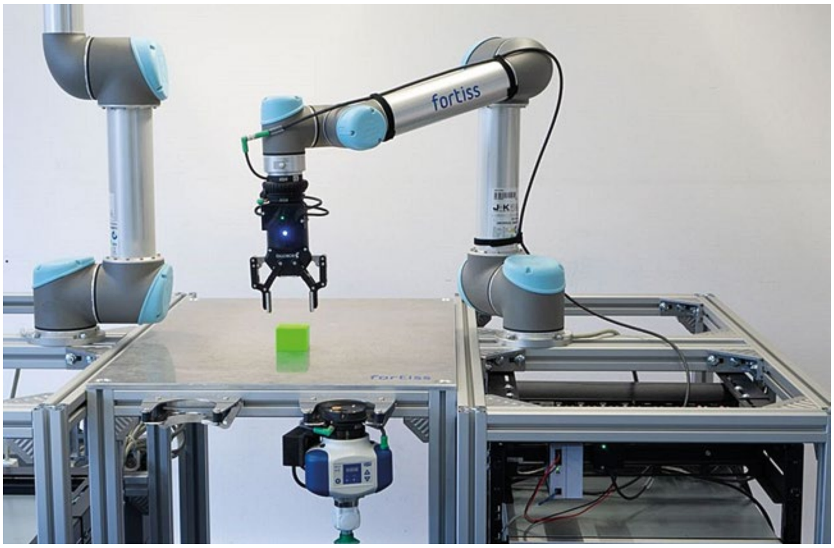


Figure 23.
Automated configuration of robot systems & analytics

Figure 24.
Knowledge-based autonomous production

combination of sensor data and semantic context information, such as from the current product, process, and manufacturing resources. Through the automatic configuration of the production system and its analytics pipeline, anomaly detection during production becomes a viable option for small batch manufacturing. The combination of symbolic knowledge and subsymbolic machine-learning approaches further leads to explainable diagnosis and the development of automatic recovery strategies.

Knowledge-based autonomous production (Figure 24) is aimed at the semantic integration and interpretation of heterogeneous data along the manufacturing company's value chains in order to enable a higher level of autonomy in the production line. These companies often struggle with handling and integrating various sources of engineering information due to a myriad of different data formats and partially non-digitalized information. In addition, the programming of complex robot systems for manufacturing tasks is time-consuming and only viable for higher production volumes. This demonstrator showcases the semantic and digital integration of heterogeneous production-relevant data across different engineering domains, from product and process design to production system engineering and the actual production. The integrated view of the relevant engineering data and its semantic interpretation enables the automatic synthesis and deployment of robot programs based on product and process specifications.

The intuitive instruction of robot systems (Figure 25) demonstrates novel concepts for SME-suitable programming of complex robot systems without the need for expert robotics knowledge. The total cost of ownership of human-robot-collaboration workcells is dominated by the operational costs for setting up and programming the robot system (60%).

Figure 25.
Intuitive instruction of robot systems

In the European Union, 99% of all manufacturing companies are small-to-medium sized enterprises (companies with fewer than 250 employees), which often lack the required expertise in robotics and industrial automation. Consequently, adoption of robot-based automation is hindered by these issues.

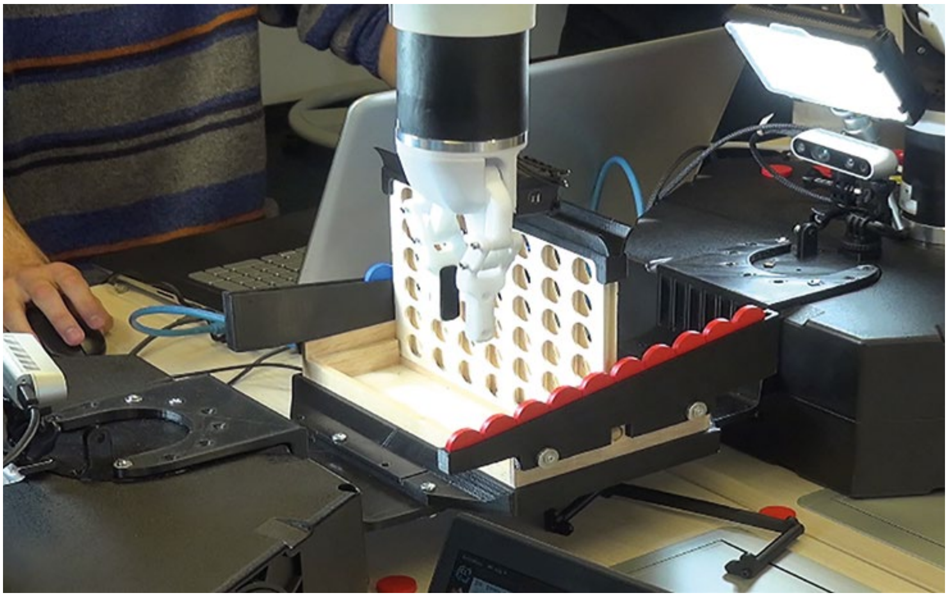
The demonstrator highlights how the combination of formal knowledge representation techniques with declarative domain-specific user interfaces enables application domain experts to intuitively formalize process descriptions in their familiar language. The higher-level and potentially underspecified user instructions can be combined with already available knowledge on the automation and application domain, in order to transform them into fully parameterized programs for robot systems.

Evaluations with real-world use cases from the mechanical engineering and woodworking domains have shown a significant increase in the efficiency of instructing robot systems, while the complexity of the human-machine interaction was reduced.

The lab course "Cognitive Robotics" (Figure 26) offers hands-on experience to bachelor's and master's students at the Technical University Munich. Participants are introduced to challenging tasks, in order to develop their own AI-enabled robot behaviors. Through weekly assignments the robot platform is gradually extended with various robot and AI functionalities, such as visual object recognition, robot control, grasp planning, world modeling and decision making.

All advances are focused on solving complex and entertaining challenges that must be tackled by interdisciplinary teams of students in a competitive manner. At the end of the lab course, tournaments based on the Connect Four and Towers of Hanoi games are held in which each team must enable their robot system to be either faster or better than their opponent. This entertaining approach to learning robotics and AI algorithms is stimulating and motivating and highly valued by the students.

Figure 26. Coginitive Robotics lab course at TUM



3.1.3 Energy Lab

Authors:
Dr. Denis Bytschkow, Dr. Markus Duchon

The fortiss Energy Lab, demonstrates research results and challenges related to current and future developments in the field of energy systems. The available demonstrators deal with various application cases and show how real problems can be addressed with findings from science and research and allow them to be presented to a wide audience in an easy-to-understand and vivid way The demonstrators are continuously developed and adapted to current problems. With our lab environment we address existing challenges in the context of research and industrial projects or in the form of student and scientific work.

In the fortiss Energy Lab, we work on topics such as the modeling of software systems, into the physical aspects. Here, we investigate opportunities and methods for designing and modeling complex systems as a basis for optimization, monitoring and control, in which the modeling of physical context and prognosis techniques from the area ML/KI are applied. Another focus area deals with evaluating system behaviors and optimizing energy systems. By using our co-simulation environment, hardware in the loop experiments can be conducted and the interactions and control mechanisms of cooperating systems can be analysed and evaluated (Bytschkow et al. 2019). The lab currently consists of the following demonstrators:

- Energy Living Lab—iEMS—software to monitor and control smart grid nodes
- Smart Electro-thermal storage
- Solar Box based on iEMS
- Energy Table—co-simulation environment

The Energy Living Lab represents an active node in the so-called smart grid. The aim of the demonstrator is to set up a laboratory-scale microgrid, networking components such as photovoltaics, battery storage, controllable loads, variable feeders, smart meters plus the components already available in the building such as the air conditioning system and the components distributed in the building, such as energy monitoring and radio-controlled sockets. This demonstrator was used to prove the technical feasibility of the approach in prototype form and to bring the possibilities of such an ICT architecture to life and test them under realistic conditions. In combination with suitable ICT networks and platforms, the protocols and gateways, application platforms and applications can be developed, tested and demonstrated under realistic conditions. In addition, the intelligent energy management system (iEMS) developed by fortiss (Duchon et al. 2014) serves as a platform for developing and evaluating different control mechanisms and systems (Rottondi et al. 2015), and for creating energy generation (Bajpai and Duchon, 2019) and load profiles, such as with the help of machine learning approaches or for connecting the entire laboratory to our co-simulation environment.

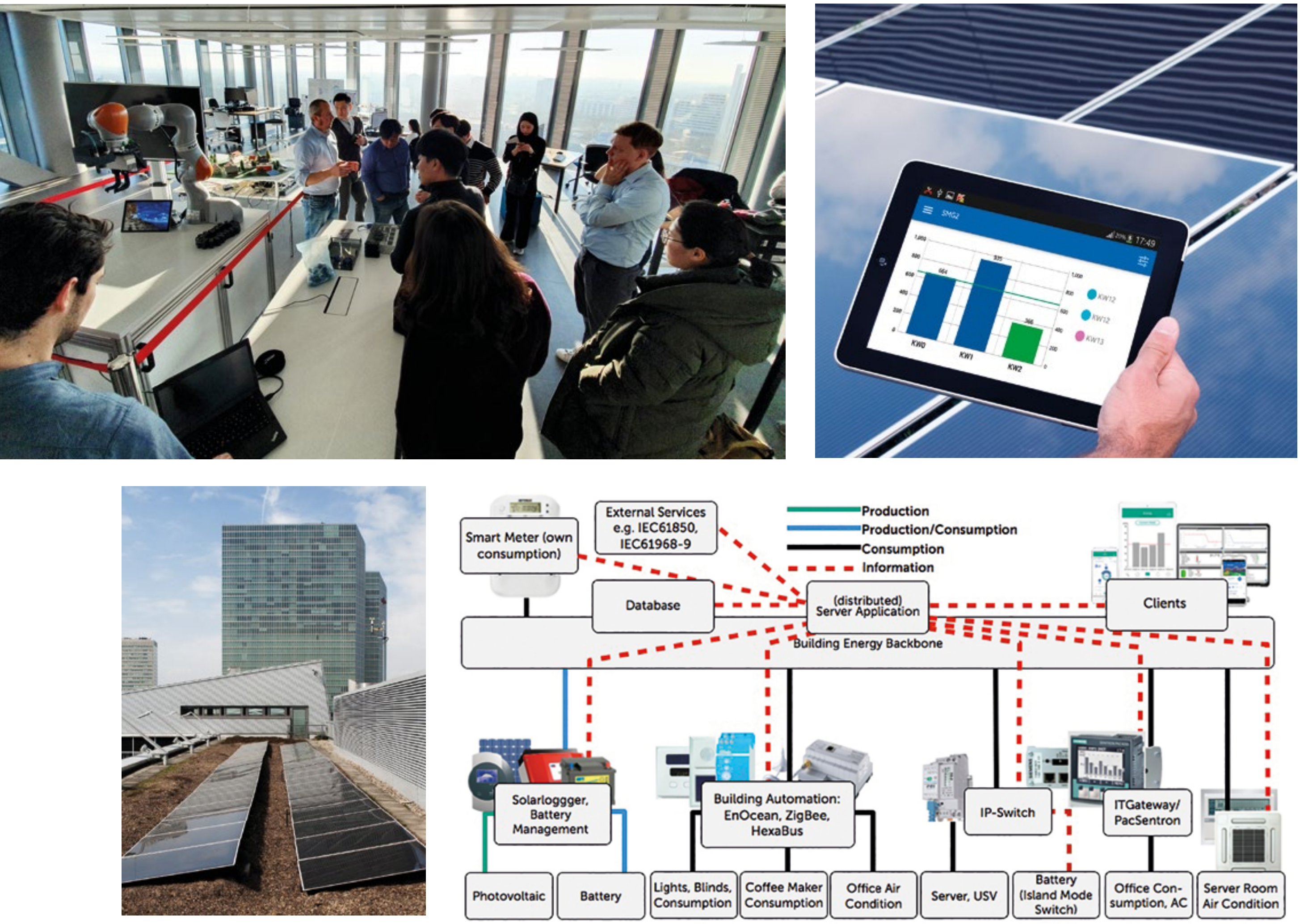

Smart Meter (own consumption)
External Services e.g. IEC61850, IEC61968-9
Production
Production/Consumption
Consumption
Information
Database
(distributed) Server Application
Clients
Building Energy Backbone
Solarloggger, Battery Management
Building Automation: EnOcean, ZigBee, HexaBus
IP-Switch
ITGateway/ PacSentron
Photovoltaic
Battery
Lights, Blinds, Consumption
Coffee Maker Consumption
Office Air Condition
Server, USV
Battery (Island Mode Switch)
Office Consumption, AC
Server Room Air Condition

Figure 27.
Energy living lab

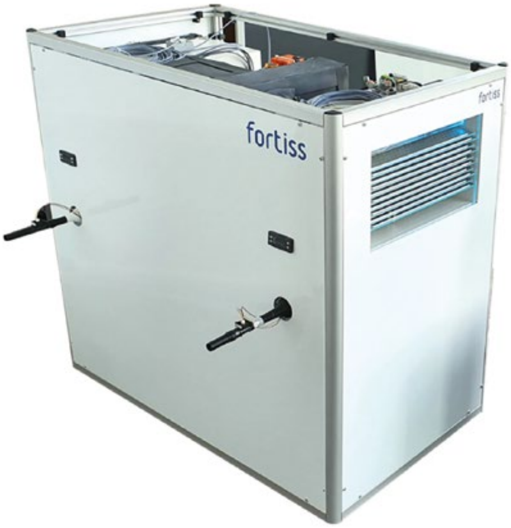
Our Smart Electric Thermal Storage converts excess electricity into thermal energy to shift energy usage in the time and energy domain. The generated thermal energy is currently stored in a phase change material (PCM) using an air ventilation system for the thermal transport to the PCM plates for heat and cold. Because this process is extremely difficult to model[120], 11 temperature sensors monitor the charging and discharging process. With this time-series data, a LSTM model[121] learns the SETS's behavior and is able to predict its state-of-charge (SOC) for the next hours. Currently, we use this demonstrator as an architecture for perpetual learning. With the adaptive method new data is periodically sent to the cloud and used to retrain the model. Once the retrained model is available the parameters are instantly updated during runtime.

Furthermore, the demonstrator implements a simple trading agent which connects the demonstrator to a virtual energy market for heat and electricity from an ongoing research project (DECENT: FKZ 0350024B). At this market the agent is able to place orders to purchase electricity when prices are low. The purchased electricity is used to generate heat and cold and is sold on the corresponding market when prices are high.

The Smart Solar Box is an intelligent solar power generator with monitoring and control capabilities. As technology continues to advance, access to

[120] J. Vogel, A. Thess, Applied Thermal Engineering, 148), pp 147–159 (2018)

[121] S. Hochreiter and J. Schmidhuber, Neural computation 9.8, pp. 1735 (1997)



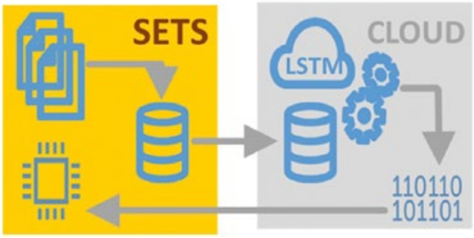


Figure 28.
Smart electric thermal storage (sets) –autonomous, ai-based sector coupler

Figure 29.
Solar box in india, workshop, and fortiss box

green and renewable energy such as solar energy is quickly gaining popularity. This is because green energy is more reliable and cleaner than most of the other available power sources. The functionality offered by this box is provided by the same iEMS that monitors and controls the Smart Energy Living lab. Here we used a limited range of functions as we only need to observe the state of charge, the consumption and the current generation.

One copy of Smart Solar Box and the fortiss software is running in a school in India and powers light and fans. In addition we carried out a workshop for

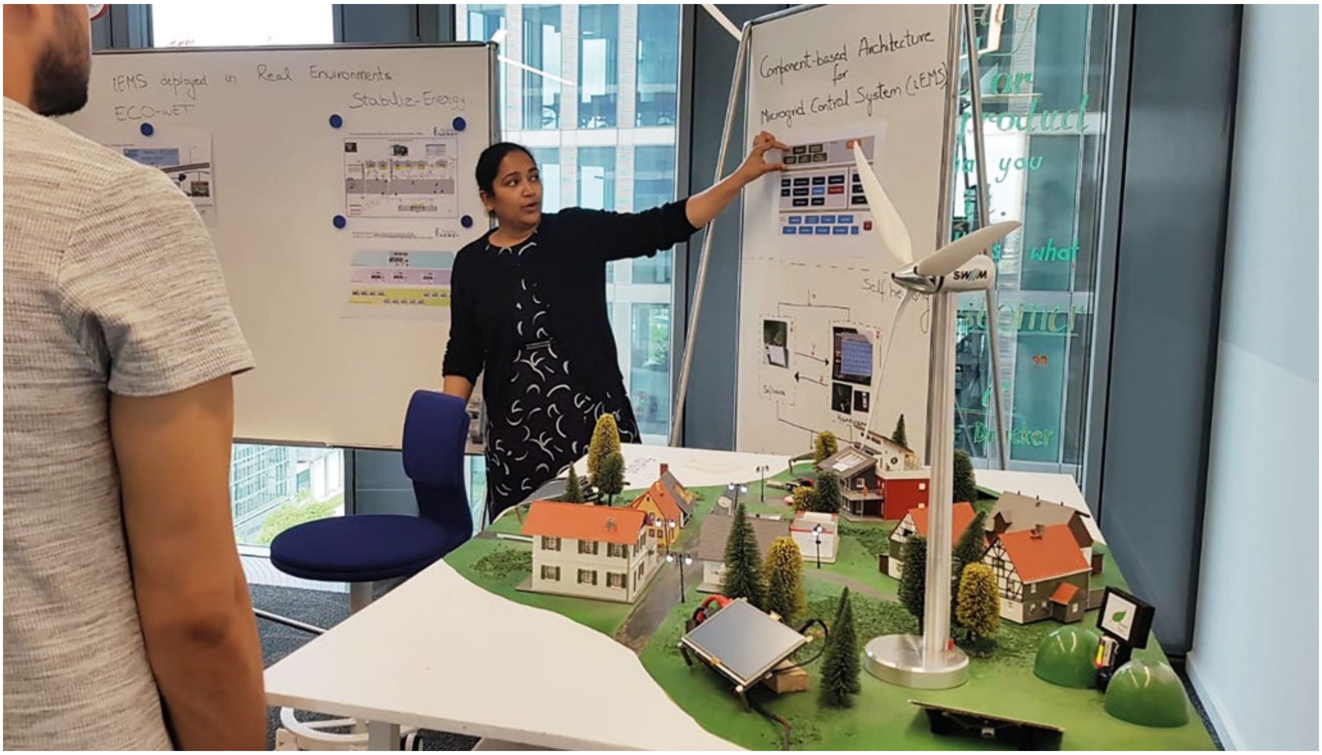

IEMS deployed in Real Environments
ECO-WET
Stabiliz-Energy
Component-based Architecture for Microgrid Control System (iEMS)
SWIM

Figure 30. Energy table

electrical engineering and computer science students where they able to build a device, deploy the software and develop additional application scenarios for decentralized energy systems.

The Energy Table integrates different software technologies developed at fortiss. It represents a small village with agricultural, commercial and residential areas. These entities are modelled with our co-simulation environment[122] (Bytschkow et al. 2015) and consume electricity according to appropriate standard load profiles. Apart from these buildings, a smart home with photovoltaic and battery storage which runs iEMS is included. It provides demand response interfaces which can be accessed by a network operator in order to disconnect the smart home from the main grid. In disconnected mode, iEMS executes degradation strategies (Gupta et al. 2015) to operate on battery supply as long as possible. In addition, another instance of the co-simulation environment represents a virtual power plant, with simulated wind turbines, biogas plant and photovoltaic systems. With the help of the controllable biogas plant, the production fluctuations of the volatile generators are balanced to cover the current demand of all buildings. Furthermore, the grid operator has two further options for balancing. During overgeneration an electric vehicle charging station can be controlled and, the streetlight system can be partly controlled only at night.

With this environment, we can showcase the functionality of a smart home controlled by iEMS, model different scenarios (consumption, generation) using the co-simulation environment, and can develop and evaluate various control strategies for distribution system operators.

[122] https://github.com/SES-fortiss/SmartGridCoSimulation

LITERATURE

Bytschkow, D., Capone, A., Mayer, J., Kramer, M., & Licklederer, T. (2019). An OPC UA-based Energy Management Platform for Multi-Energy Prosumers in Districts. In Proceedings of 2019 IEEE PES Innovative Smart Grid Technologies Europe, ISGT-Europe 2019.

Duchon, M., Gupta, P. K., Koss, D., Bytschkow, D., Schätz, B., & Wilzbach, S. (2014). Advancement of a sensor aided smart grid node architecture. In Proceedings - 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2014 (pp. 349–356).

Bajpai, A., & Duchon, M. (2019). A hybrid approach of solar power forecasting using machine learning. 2019 3rd International Conference on Smart Grid and Smart Cities (ICSGSC), 108–113.

Bytschkow, D., Zellner, M., & Duchon, M. (2015). Combining SCADA, CIM, GridLab-D and AKKA for smart grid co-simulation. In 2015 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2015

Gupta, P. K., Becker, K., Duchon, M., & Schatz, B. (2015). Formalizing performance degradation strategies as an enabler for self-healing smart energy systems. In Tagungsband - Dagstuhl-Workshop MBEES 2015: Modellbasierte Entwicklung Eingebetteter Systeme XI (pp. 110–119).

Rottondi, C., Duchon, M., Koss, D., Palamarciuc, A., Pití, A., Verticale, G., & Schätz, B. (2015). An energy management service for the smart office. Energies, 8(10), 11667–11684.

3.1.4 Mobility Lab

Authors:
Simon Barner, Tobias Kessler

Engineering of cyber-physical systems such as autonomous cars is extremely challenging. This is due not only to the complexity of ADAS functions and the hardware/software platforms that provide the required performance, but also the need to deliver new or updated advanced software-defined functions over the entire life cycle of a car. In the fortissimo Rover Model-based Systems Engineering Lab, we investigate how model-based systems engineering can be employed to tackle these challenges. We focus on platooning, an autonomous driving function, which permits automobiles or trucks to drive behind one another at extremely close distances, thus reducing fuel consumption.

The lab is our focal point and figurehead for research on model-based systems engineering methods, languages and tools for cyber-physical systems, where we currently examine the following research issues:

- Model-based method for deriving assurance cases (Cârlan et. al, 2017; Cârlan et al., 2019) for validating the functional safety of the vehicles in line with ISO-26262.
- Degradation and reconfiguration strategies (Becker et al., 2018) for safeguarding critical driving functions (e.g., against hardware faults).
- Analytical and simulation-based processes for dimensioning and validating the vehicle hardware and software architecture (Eder et al., 2020).
- Co-simulation of cyber-physical systems: ADAS functions, vehicle dynamics, fault injection.
- Defect-based integration testing for CPS: elicitation and operationalizing of defect models.

The lab currently features a demonstrator based on the fortissimo platform, including a co-simulation environment, as well as a real-world prototype of an autonomous vehicle that we call fortuna.

Using the fortissimo platform, we conduct research into platooning components such as joining or leaving a platoon and car-to-car communication, as well as ADAS functions like adaptive cruise control systems and lane/emergency braking assistants. These functions are implemented in terms of behavior models in the AutoFOCUS3 open source systems engineering tool (fortiss GmbH; Aravantinos et al., 2015) developed by fortiss. An initial validation of the models can be performed in a functional simulation environment (Lúcio et al., 2018a) and is

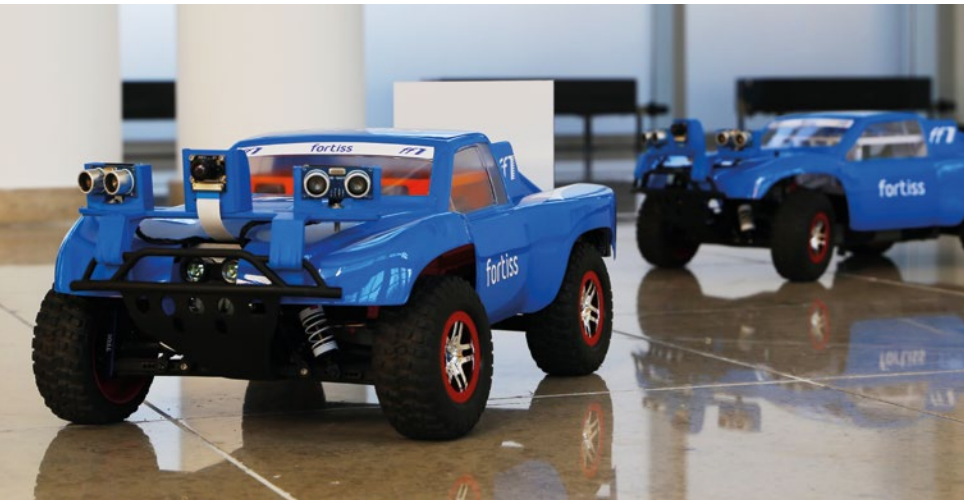


Figure 31.
Fortissimo rover hardware platform – model vehicle equipped with 3D-printed components, raspberry pi controllers and sensor technology.

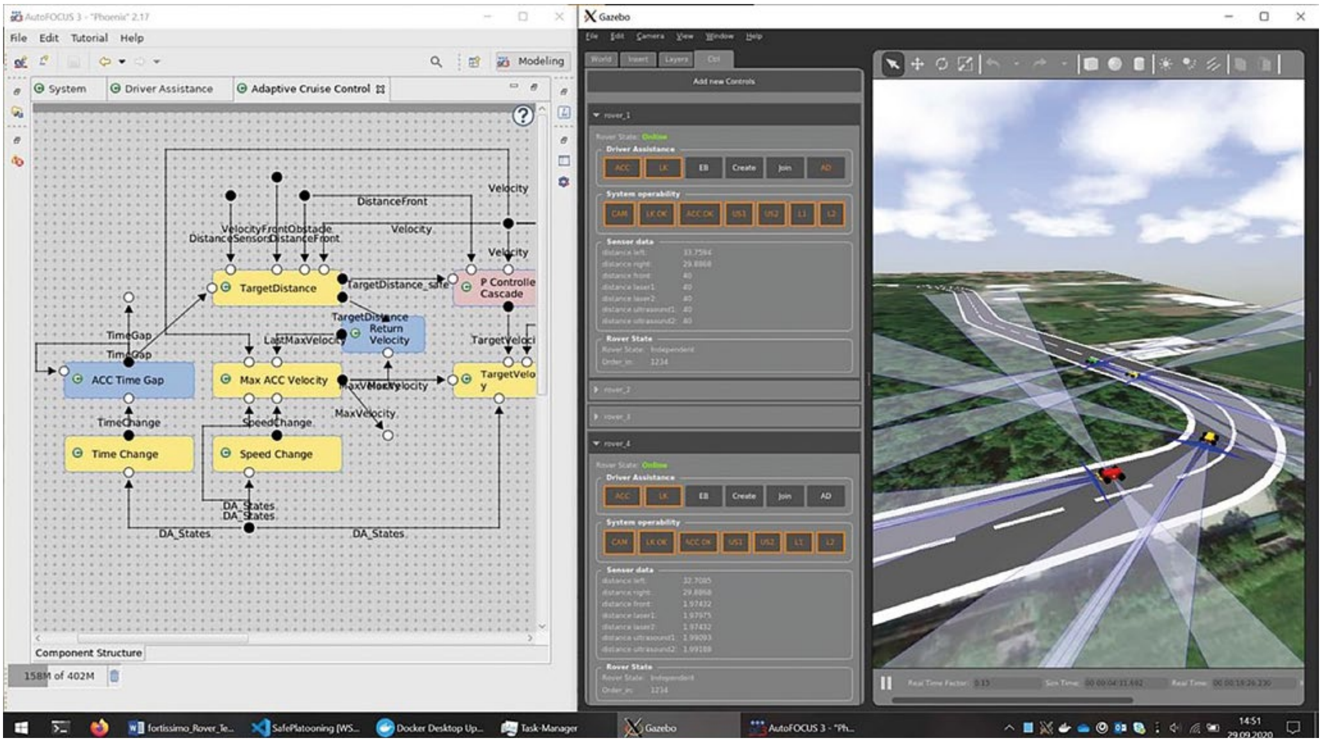


Figure 32.
Virtual prototype of the fortissimo rover: co-simulation of the vehicle functions, physical behavior and environment.

supported by a co-simulation of the dynamics of the vehicle and its environment in other open source tools such as OpenModelica, ROS and Gazebo, which also provides a 3D-simulation. As a final step, the developed functions are implemented via code generators in the fortissimo Rovers, 1:10 scale model vehicles equipped with sensor technology such as cameras, ultrasound and laser-based distance meters (see Figure 31).

The fortissimo Lab also serves as basis for a bachelor's/master's practical course titled "From sensors to driving functions—develop your own car", which the Model-based Systems Engineering field of competence at fortiss regularly organizes together with the chair for Software & Systems Engineering at Technical University Munich. Moreover, weare planning to offer training programs for model-based systems engineering (e.g., based on the AutoFOCUS3 tutorial presented at the MODELs conference (Lúcio et al., 2018b) and the online courses developed in the SPEDiT research project (SPEDiT consortium).

fortuna is designed as a street-legal, full-size demonstrator of a cyber-physical autonomous system. The basis for the platform is a retrofitted VW Passat GTE Plugin Hybrid vehicle (see Figure 33). Autonomous driving is a highly vivid research area. Many industry players in the field have announced market-ready systems more than once. To date however, there are no vehicles with correct autonomous driving functionality sharing the roads with human drivers. A main barrier towards market-ready fully autonomous cars is the basic difficulty of transferring the performance from research, predevelopment and simulated systems to real roads. To show the applicability of our research on joint action planning in real-world scenarios, we maintain a vehicle equipped for fully autonomous driving.

With the comprehensive sensor setup, our car is capable of evaluating the state-of-the-art perception and sensor data fusion approaches and is equipped with a state-of-the-art sensor set, such as a 360° Lidar setup for a high-quality,

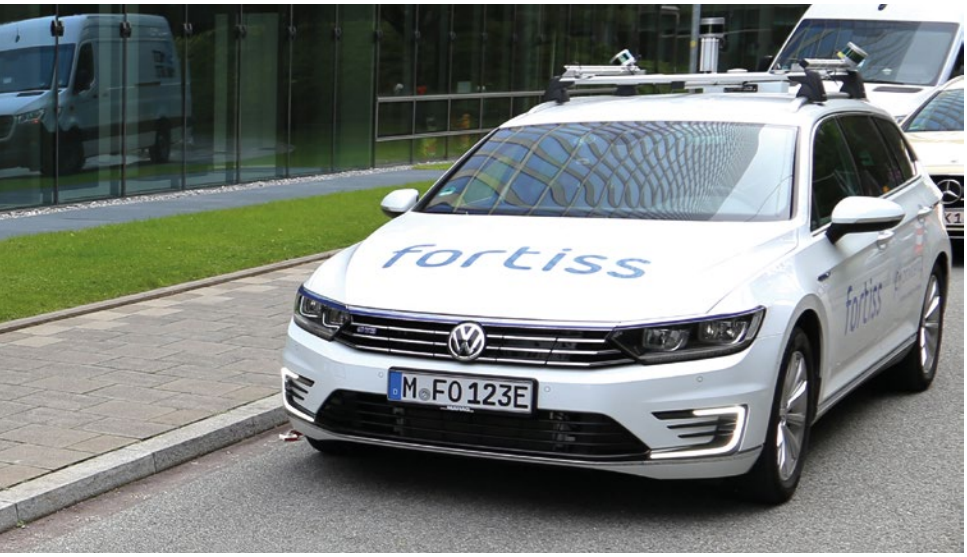


Figure 33.
The autonomous driving prototype vehicle fortuna.

accurate 3d representation of the environment around the vehicle. Within the research project *Providentia* the sensor setup is extended even more. fortuna is used as a sensor platform to demonstrate the capabilities of a digital twin build from infrastructure sensors (Krämmer, A. et al., 2019).

Furthermore, the platform offers all necessary interfaces and computing power for operating in fully-automated mode. Given that autonomous driving applications require an enormous amount of software, we focus on our core expertise such as the development of joint action planning modules (cf. 2.3—Joint Action Planning) and we base the software stack on the well-known apollo open-source project, to which we contributed.[123] The methodology has also been made publically available (Kessler, T et al., 2019, pp. 1612–1619). Basing the software stack on apollo makes a performance baseline available and helps to track the software improvements. As the demonstrator is street-legal in Germany, we can demonstrate the applicability of research and experience the challenging scenarios in a real environment.

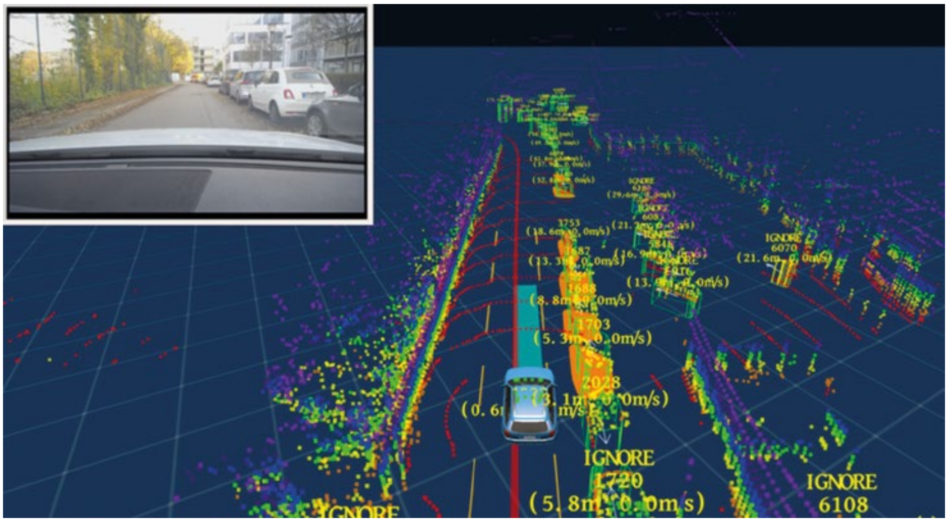


Figure 34.
Raw and processed sensor data as seen by the vehicle. The trajectory of the vehicle is depicted in blue.

[123] https://github.com/fortiss/apollo

LITERATURE

Aravantinos, V., Voss, S., Teufl, S., Hölzl, F., & Schätz, B. (2015). AutoFOCUS 3: Tooling Concepts for Seamless, Model-based Development of Embedded Systems. Proc. 8th Int. Workshop Model-based Architecting of Cyber-Physical and Embedded Systems (ACES-MB). 19–26.

Becker, K., Voss, S., & Schätz, B. (2018). Formal analysis of feature degradation in fault-tolerant automotive systems. Science of Computer Programming, 154, 89–133. doi: 10.1016/j.scico.2017.10.007.

Cârlan, C., Barner, S., Diewald, A., Tsalidis, A., & S. Voss (2017). ExplicitCase: Integrated Model-based Development of System and Safety Cases. Proceedings of the SAFECOMP 2017 Workshops ASSURE, DECSoS, SASSUR, TELERISE, and TIPS, 10489, 52–63. doi: 10.1007/978-3-319-66284-8_5.

Cârlan, C., Nigam, V., Voss, S., & Tsalidis, A. (2019). ExplicitCase: Tool-Support for Creating and Maintaining Assurance Arguments Integrated with System Model. 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 330–337. doi: 10.1109/ISSREW.2019.00093.

Eder, J., Voss, S., Bayha, A., Ipatiov, A., & Khalil, M. (2020). Hardware architecture exploration: automatic exploration of distributed automotive hardware architectures. Software and Systems Modeling, 19, 911–934. doi: 10.1007/s10270-020-00786-6.

fortiss GmbH. AutoFOCUS3—Model-based development of embedded systems. Retrieved Feb. 1st, 2021, from https://www.fortiss.org/en/publications/software/autofocus-3.

Lúcio, L., Kanav, S., Bayha, A., & Eder, J. (2018a). Controlling a virtual rover using AutoFOCUS3. Proceedings of the MDETools Workshop co-located with MODELS 2018, 2245, 356–365. http://ceur-ws.org/Vol-2245/mdetools_paper_6.pdf.

Lúcio, L., Voss, S., Chuprina, T., Bayha, A., Eder, J., & Kanav, S. (2018b). [T3] Develop your Own Car, MODELS Conference Tutorials. Copenhagen, Denmark. https://modelsconf2018.github.io/program/tutorials/#t3-develop-your-own-car.

Krämmer, A. et al. (2019) Providentia—A Large Scale Sensing System for the Assistance of Autonomous Vehicles, Robotics Science and Systems Workshops (RSS Workshops).

Kessler, T. et al. (2019). Bridging the Gap between Open Source Software and Vehicle Hardware for Autonomous Driving. 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, pp. 1612–1619, doi: 10.1109/IVS.2019.8813784.

SPEDiT consortium. Training materials for model-based development of embedded systems. Retrieved Feb. 1st, 2021, from https://cce.fortiss.org/spedit/.

3.1.5 Drone Lab

Authors:
Dr. Ernest Wozniak, Florian Grötzner

Unmanned aerial vehicles, commonly known as drones, is a solid industry with a wdie range of applications that demonstrate their usefulness. While remotely controlled drones have been marketed for some time now, their autonomous operation still requires advanced research in order to support industry's goal in this direction. This relates especially to drones designed for safety-critical applications, such as taxis or drones that operate in urban areas.

In a simplistic view, a trustworthy operation of an autonomous drone may be largely assured by focusing on two aspects: first a guarantee of the trustworthiness of AI components, and second, the correct operation of the integrated system under the assumption that trustworthiness guarantees of each component were achieved.

The fortiss Drone Lab aims to support the industrial need for innovative functionality that requires trustworthy autonomous drones. This requires targeting the problem from the two previously-mentioned perspectives. From the holistic (integrated) system perspective, the fortiss Drone Lab focuses on the following goals:

- Providing the ability to test new autonomous systems (drones) in a simulated environment. It should then be possible to directly transfer the software to a real-world drone.
- Developing a modular architecture for the simulator and the autonomous drones, in order to allow for rapid testing of new functionality (such as algorithms, sensors, or drone type).
- Designing the behavior of an autonomous drone with a clear differentiation between distinctive levels of behavior (including the modular structure of components). This is useful for a clear definition of responsibilities but also for the activities related to certification or trustworthiness analysis.
- Possibility to test a hardware platform in a simulated environment that enables the seamless transition of an autonomous behavior software stack.

Fostering single components that contribute to the autonomous behavior of a drone results in other issues that must be examined:

- Focusing on high-level intelligent behavior and using well-established technology for low-level functionality such as standard control software.
- Enabling a drone to cope with unknown conditions through insertion of human/expert knowledge into suitable AI/ML components.

The first three objectives of the integrated system perspective were reached. We developed a simulation platform to test autonomous drones. The platform uses the open source flight-control stack PX4[124], which comes with pre-implemented standard functionality such as hovering or flying to checkpoints. The objective of modularity has been achieved by using a Gazebo simulation that integrates different drones, sensors and algorithms (see Figure 35).

[124] PX4 Autopilot, URI: https://docs.px4.io/master/en/

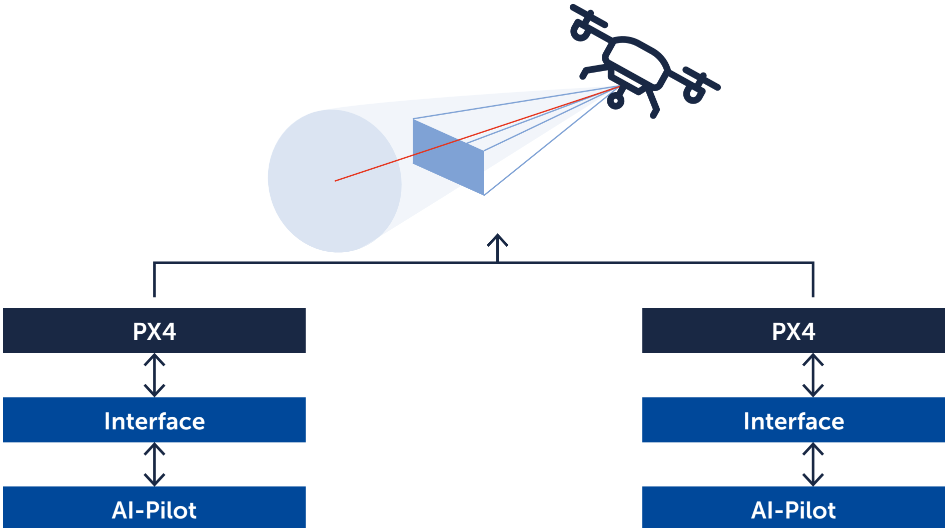
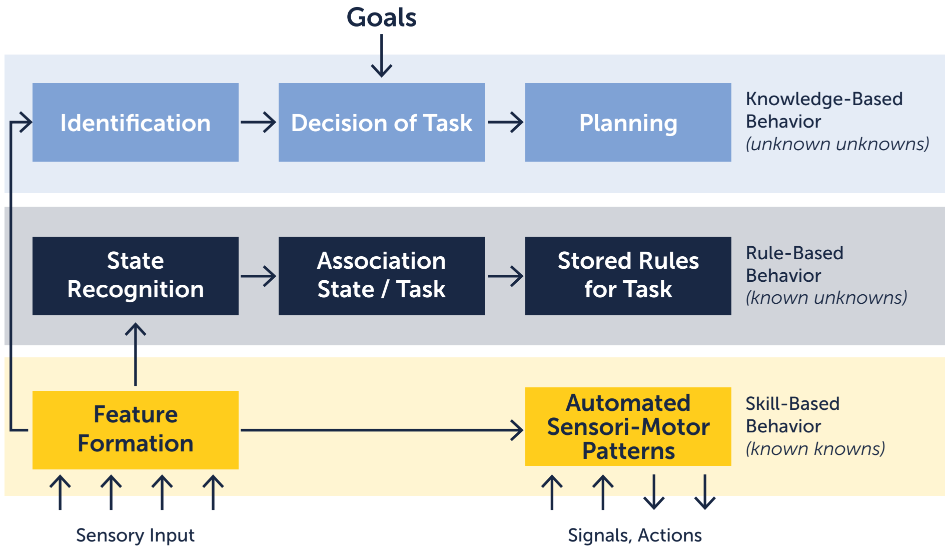


Figure 35.
Gazebo-based simulation of drones controlled by PX4

Further, we have developed an AI pilot concept that constitutes a module for high-level intelligent behavior. In order to concretize it, we specified a solution-level white box model (refer to VDE standard part 3—VDE, 2020 for details on solution-level models) with a clear differentiation between distinctive levels of behavior. It is based on the Rasmussen[125] [126] model (see Figure 36) for cognitive processes, and it decouples functionalities contributing to autonomous behavior generation in a horizontal and vertical manner. Horizontal decomposition identifies sensing and perception (left column), decision on new tasks or task selection (middle column), and execution (right column). The vertical decomposition separates functionalities into three layers: skill-, rule- and knowledge based behavior. These layers are responsible for performing tasks of increasing complexity and as a result require a greater level of understanding and knowledge about the situation that the autonomous drone is facing. In order to reach the objective of

Figure 36.
The Rasmussen Model of Cognition



[125] Rasmussen, J. (1985). The role of hierarchical knowledge representation in decisionmaking and system management. IEEE Transactions on systems, man, and cybernetics, (2), 234–243.

[126] Rasmussen, J. (1987). Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering.

a seamless transition from simulation to real-world scenarios, we are planning to implement an autonomous flight stack based on the Rasmussen scheme, compliant with the simulation engine.

One current activity mainly concerns fostering single components that contribute to the autonomous behavior of a drone, such as targeting the last two topics from the list of predefined goals. A relevant aspect of these two topics is transitioning from rule to the knowledge-based behavior (also called "known unknowns" and "unknown unknowns") in the Rasmussen model. The "unknown-unknowns" require high-level and intelligent reasoning, similar to that of a human being, which has the capability of inferring tasks and solutions by a retrospective consideration of its experience and acquired knowledge.

A very promising research line is the explicit integration of additional knowledge into AI components (such as those built with deep neural networks). Although knowledge representations (ontologies) and their integration into deep neural networks can be domain-independent, the type of knowledge useful for drones in order to cope with "unknown-unknown" scenarios is different due to the different nature of unknown-unknowns. By relying on a simulation environment, one can test knowledge-enhanced components by crafting scenarios not experienced by a drone in advance.

LITERATURE

VDE-AR-E 2842-61 - Design and Trustworthiness of autonomous/cognitive systems, 2020.

3.2
IBM fortiss Center for AI

Author:
Dr. Holger Pfeifer

IBM and fortiss founded a joint research Center for Artificial Intelligence (C4AI), which is colocated at the IBM Watson Center Munich and aims to create innovative, reliable and secure AI technologies for business and society. The IBM fortiss Center for AI is globally networked with research and application partners from Germany, Switzerland, Ireland, and the US. In the joint facility, around 40 scientists research and develop new AI-supported software solutions for mission and business critical applications, both for industry and the public sector. In cooperation with our partner IBM we are successively building up a portfolio of solution-oriented AI projects to sustainably tap the potential of AI. Current focus is placed on: AI-based assistance systems for transparent and human-centered decision support, intelligent control of autonomous robots, prototyping of new digital citizen services, and trustworthy and privacy-preserving machine learning. New projects are defined in an agile manner when new business needs or opportunities are identified.

3.2.1. Accountable Federated Machine Learning

Authors:
Dian Balta, Dr. habil. Ulrich Schöpp, Mahdi Sellami

Sharing knowledge without releasing data? This question is posed by numerous actors, who would like to benefit from machine learning developments but are not able to share the required data due to regulatory, legal or business restrictions. The question is how a consortium of actors can curate knowledge from distributed data, if the actors cannot share the data for learning purposes? This question arises in several domains, such as banking, healthcare and public administration. Additionally, an answer to the question should involve generating verifiable claims along the learning process, since potential applications often underlie strict laws and regulations such as GDPR and actors are publicly accountable for their actions. Through publications, software modules and demonstration prototypes, we addressed this question by augmenting the applicability of federated machine learning in a civic participation case.

Our research exemplifies for the very first time, how knowledge curation based on advanced machine learning techniques can be achieved in a federated setup where trust is built upon auditable claims and tamper-proof evidence on compliance with rules. In terms of implications, we describe how real-world applications in federal settings such as the German Federal Government system can benefit from machine learning while being legally compliant.

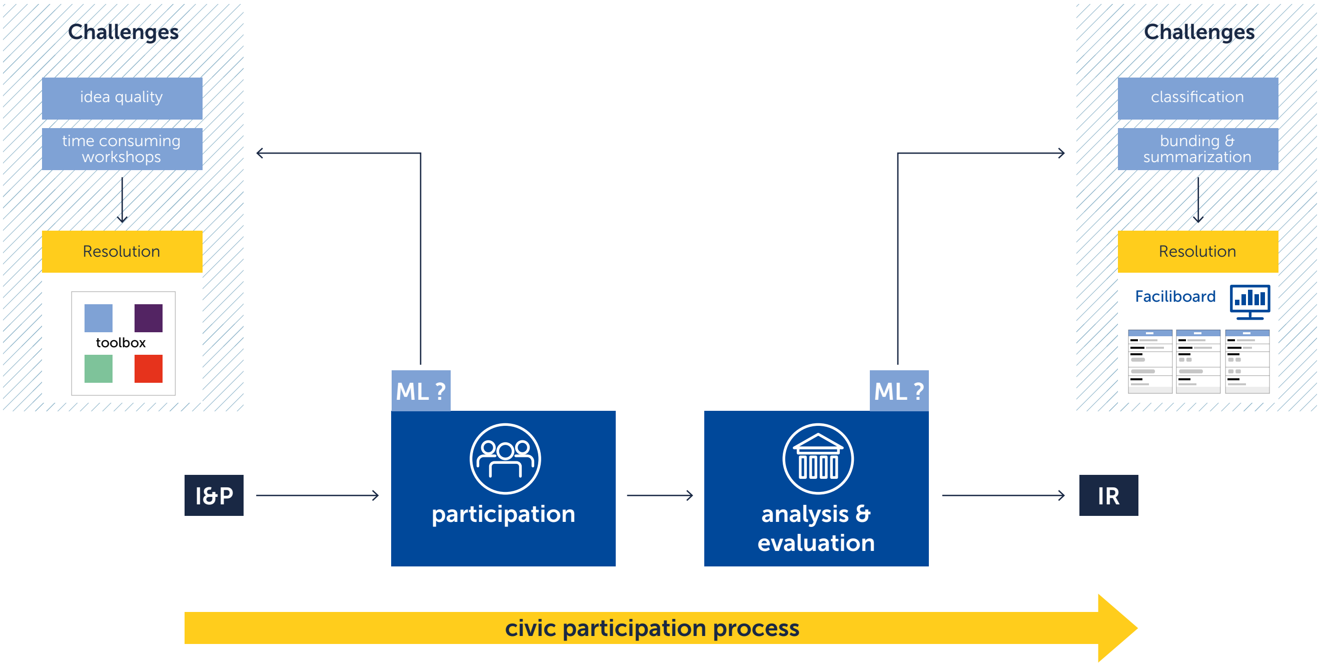


Figure 37.
Challenges addressed by machine learning in civic participation

Federated machine learning in a federal government setting

Federated machine learning (FML) is an approach to allow multiple parties to cooperatively build a common machine learning model from their data without having to share this data. The idea is that all the parties execute machine learning tasks on their private data sets and exchange the resulting model updates to produce a combined model of the whole data. In this way, the data remains private and the parties exchange only model updates and testing data for assessing the quality of learned models.

Consider a system that allows citizens to make improvement suggestions for plans by the municipal administration (Balta et al, 2019). The government wants to use machine learning to classify and analyze suggestions, so that they can be processed more effectively. Having been successfully tested in one city, a number of cities decide to roll out the system as well. It would now be desirable to analyze the combined data to produce higher quality models for classifying suggestions. However, due to Germany's federal structure, particularly with respect to data privacy, the cities may not be able to share their data directly. They might prefer to use FML, perhaps combined with other privacy measures, to share only the machine-learned knowledge, but no data. The underlying learning principle here is "share knowledge, do not release data".

An important goal of federated machine learning is to produce models of *high quality* despite data not being shared. This is not an easy task, and FML therefore builds on sophisticated algorithms that require non-trivial interactions between the participating parties. This requires defining claims about the FML process that can be audited by the cities—or by the citizens. For instance, a valid claim should provide verifiable evidence that the data analysis was not biased by one civic group.

The need for accountability: factsheets, auditable claims and evidence

Mechanisms for supporting auditable claims[127] during the development of machine learning applications represent one potential approach to increasing trustworthiness in ML. Such mechanisms provide the ability to make precise claims for which evidence can be brought to bear, so that ML developers can more readily demonstrate responsible behavior to regulators, the public, and one another. The need for such mechanisms is particularly present when FML is applied, given the opacity of the learning process, available distributed data, decentralized governance of the consortium as well as the non-trivial interactions between the parties involved.

Accountability is needed in order to allow the participants as well as auditors to trust the federated machine learning process. It must be possible to implement the learning process in an accountable way, so that, despite its complexity, all participants can be convinced that it has been carried out correctly according to set rules, that all participants have been treated fairly and equally and that no participant has manipulated inputs or introduced bias for personal gain. Decisions, such as to exclude contributions by a participant to ensure the quality of the overall model, should be made objectively and participants should be able to reproduce them. If needed for an audit, results should be verified by reproduction of the learning process. To make FML processes accountable, they must be extended with documentation of all essential actions and decisions. Participants must all be able to agree that this documentation accurately represents what has been performed. It should be sufficient to establish trust in the produced model, even when participants may not fully trust one another.

Our approach is based on the idea of factsheets[128], which were proposed as a way to provide transparency and establish trust in (non-federated) AI applications. Factsheets are intended to be delivered together with AI models to provide essential information. They document what data has been used to train the model, what algorithms were used and what parameters. They document important model properties, such as performance, fairness, robustness, explainability and lineage.

Benefits: accountable federated machine learning

With AFML, we show how knowledge curation based on advanced machine learning techniques can be achieved in a federated setup where trust is built upon auditable claims and tamper-proof evidence on compliance with rules. Our approach was to extend factsheets for federated machine learning and to define a level of accountability. With respect to our research results, practitioners can design, develop and operate of machine learning applications by relying on auditable claims about their compliance.

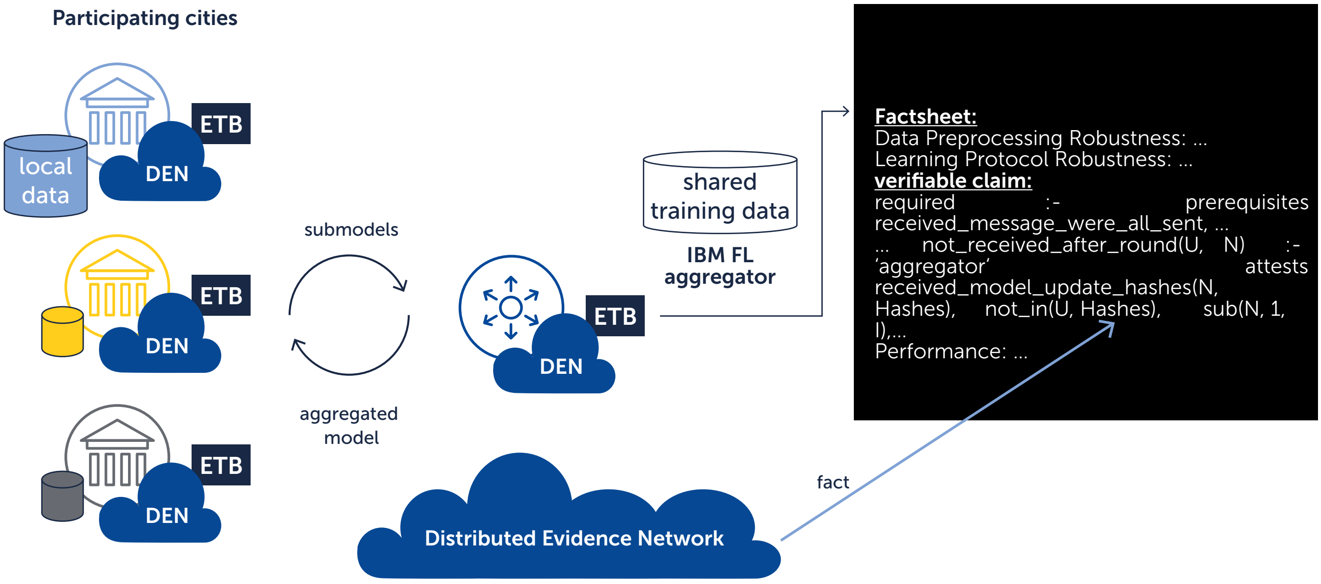
- Architecture: we described an architecture that combines FML and accountability tools on top of existing data infrastructures.
- Factsheets for FML: we defined factsheets for FML, which includes a formalized workflow that generates auditable claims by providing a formalized semantics for tamper-proof evidence.
- Demonstrator: We implemented a demonstrator using data from a real-world use case of civic participation in Germany (AFML, 2020).

Architecture: We extended the IBM Federated Learning (IBMFL) framework[129] with an accountability component.

[127] cf. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.

[128] IBM AI factsheet project, https://aifs360.mybluemix.net/

[129] Cf. https://github.com/IBM/federated-learning-lib



Factsheet:
Data Preprocessing Robustness: ...
Learning Protocol Robustness: ...
verifiable claim:
required :- prerequisites received_message_were_all_sent, ...
... not_received_after_round(U, N) :- ‘aggregator‘ attests received_model_update_hashes(N, Hashes), not_in(U, Hashes), sub(N, 1, I),...
Performance: ...

Figure 38.
Use case architecture

As a first use case, we are implementing a prototype for the citizen participation example outlined above. Despite the decentralized approach, the system is designed to ensure a traceable and verifiable process while guaranteeing that the results that are generated adhere to criteria such as data protection, security and accuracy. The concept is based on data and models in the context of civic participation that will be used to automatically group various input from citizens according to the subject or issue.

The architecture of the use case is summarized in Figure 39. It shows the computation nodes of the IBMFL framework in blue and Evidentia nodes in yellow. Each IBMFL node has an associated ETB node from the Evidentia framework, which is used for recording actions and for executing verification workflows. The ETB nodes produce a record of the learning process on a ledger, from which they can generate factsheets.

Factsheets for FML: Factsheets for federated machine learning collect information from all involved participants. They record the rules of federation that the participants agreed on, their interactions during the learning process, what has been done and by whom and what decisions were being made and why. Claims in the factsheet need to be traceable to the participants and auditors should find enough information to assess the veracity of the participants’ claims. This is not easy to achieve, as one must cover a wide range of scenarios. There are many possible trust relationships between the participants and federated machine learning workflows can differ substantially depending on the choice of algorithm and parameters.

Demonstrator: We demonstrated a prototype implementation based on the fortiss Evidentia framework (Evidentia, 2020). It builds on a distributed ledger (referred to as a distributed evidence network, or DEN) to allow a consortium of actors to record claims in an auditable manner, even in the absence of mutual trust. It uses a logical specification mechanism for formalizing accountability workflows and for collecting information into factsheets. Its flexibility allows us to integrate many services and techniques and to cover a wide range of scenarios.

At a high level, we use Evidentia to integrate the documentation tasks that are needed for accountability into the federated machine learning process. It provides the participants with the means to document their actions on a tamper-proof common record. It implements workflows (based on Datalog and through a component called evidential tools bus, or ETB (refer to ETB) to continuously verify that the record entries match the actual actions of the participants and that they conform to the greed-upon learning process. It allows the user to integrate various verification methods, from simple spot checks to fully formal proofs. It assembles the recorded information into factsheets.

Directions and challenges

Next steps would include the integration advanced methods for verifying content, in addition to processes, such as trusted execution environments and cryptographic methods. It also requires the development of new AI methods for efficiently auditing the results of machine learning algorithms. Likewise, various possible trust relationships should be captured and supported by formal security proofs of the auditing protocols. It verifies process implements trust models, which are security proofs for various trust relationships.

LITERATURE

AFML (2020): https://git.fortiss.org/c4ai-afml.

Balta, D., Kuhn, P., Sellami, M., Kulus, D., Lieven, C., & Krcmar, H. (2019). How to streamline AI application in government? A case study on citizen participation in Germany. In International Conference on Electronic Government (pp. 233–247). Springer, Cham.

ETB - Evidential Tool Bus: https://www.fortiss.org/forschung/projekte/detail/evidential-tool-bus.

Evidentia (2020): https://git.fortiss.org/evidentia.

3.2.2. Proactive & interactionless government services

Authors:
Dian Balta, Peter Kuhn

The future of public administration is proactive and interactionless. Government services shall be automatically provided without the need for applications and without the user having to interact with an application. For the provision of such proactive and interactionless services, intelligent data processing using machine learning and accountable data exchange using distributed ledger technology (DLT) will form the basis of the technology.

Our approach in this project was to apply the concept of proactive and interactionless government services to real scenarios. Therefore, we have developed and applied an analysis method for the readiness of a particular service, extended existing software frameworks and developed two demonstrators (for child benefit services as well as for applying for a restaurant license). Through our research, we offer government practitioners a structured engineering approach to link visionary service design with advanced technologies towards higher service quality for citizens and businesses.

Proactivity and non-interaction

The notion of proactivity in government has been a topic of research in the context of public services from different perspectives and for different aspects[130]. Proactive service provision by governments can be defined as delivering "a service to a citizen when a life event occurs, without the citizen having to request the service"[131]. A government that delivers proactive services is considered user-friendly and improving service quality, since it supplies a service to the user (user-centered) instead of just approving it (government-centered). Three levels of proactivity can be distinguished for governments: a reactive government that is not proactive at all, an attentive government that has some proactive aspects, and the fully proactive government that is proactive in all aspects.

In a continuous interpretation, proactivity of a service can be seen as inversely proportional to the interaction effort for the user to get the service[132]. Completely proactive services in the spirit of this interpretation are therefore non-interactive, in other words they do not require any user-government interaction.

From a user perspective, proactive governments result in reduction or complete absence of interactions to obtain a specific service. Given that interactions such as filling out and filing forms are considered cumbersome by users, their reduction or complete absence potentially has positive effects on service quality. Arguably, non-interaction can be considered a major factor determining the perception of public service and should be a focus of government efforts to increase service quality. Implementing non-interaction would require a novel design of data provision as well as supporting functions during service provision.

Proactive public administration takes matters into its own hands and helps its users in order to reduce the effort to a minimum. This is achieved through the use of uniform interfaces between the participating IT systems, intelligent data processing using ML and accountable data exchange according to regulations and legal constraints using DLT.

[130] Linders, D., Liao, C.Z.-P.,Wang, C.-M.: Proactive e-Governance: Flipping the service delivery model from pull to push in Taiwan. Govern. Inf. Q. 35, 68–76 (2018).

[131] Scholta, H., Mertens, W., Kowalkiewicz, M., Becker, J.: From one-stop shop to no-stop shop: An e-government stage model. Govern. Inf. Q. 36, 11–26 (2019)

[132] Brüggemeier, M.: Auf dem Weg zur No-Stop-Verwaltung. Verwaltung Management 16, 93–101 (2010).

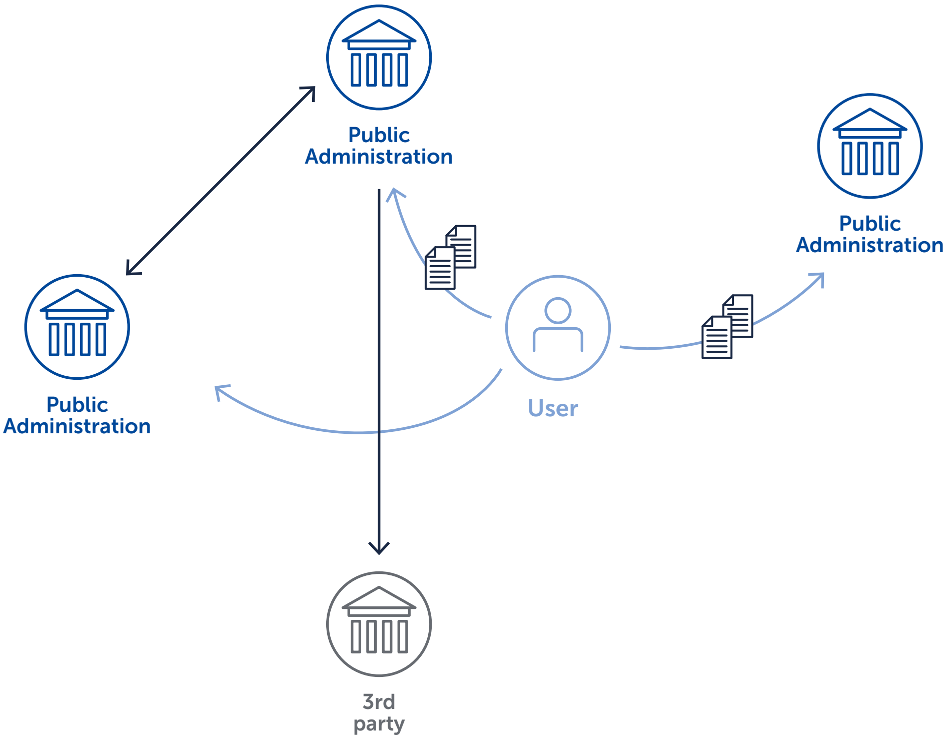


Figure 39.
User-centric Service

ML and DLT for government services

One application of ML with particular relevance for government services is a methodology referred to as natural language processing (NLP). With NLP, tasks such as information extraction and summarization or discourse and dialogue or even machine translation can be automated to a certain degree. Consequently, the goals of applying NLP in online citizen participation include designing a more efficient process by supporting the ideation (suggesting key-words or related contributions during ideation) as well as the analysis and evaluation (clustering and classifying user contributions). NLP has been previously employed in government applications[133]. While various tools and automated programmable interfaces (APIs) exist, recent analysis shows that open source tools, which allow for better control of data privacy and on-premise operation of NLP, perform well with established API providers compared to closed source logical interference software and knowledge models.

DLT, including but not limited to blockchain, is a combination of well-known computer science, cryptography and economic concepts: linked lists, distributed networking, hashing, digital signatures, asymmetric encryption, ledgers and incentive mechanisms for coordination of participants towards building consensus[134]. It allows "secure processing of transactions between untrustworthy parties in a decentralized system", while maintaining a single point of truth. A smart contract is a computer program which can be used by the participants inside a DLT network. It automatically executes transactional events, if pre-specified contractual terms are fulfilled and can be used to avoided manual document checking for instance.

[133] Cf. e.g. Androutsopoulou, A., Karacapilidis, N., Loukis, E., Charalabidis, Y.: Transforming the communication between citizens and government through AI-guided chatbots. Government Information Quarterly. (2018).

[134] cf. e.g. Buchinger, M., Balta, D., & Krcmar, H. Distributed Ledger Technology in the Banking Sector: A Method for the Evaluation of Use Cases.

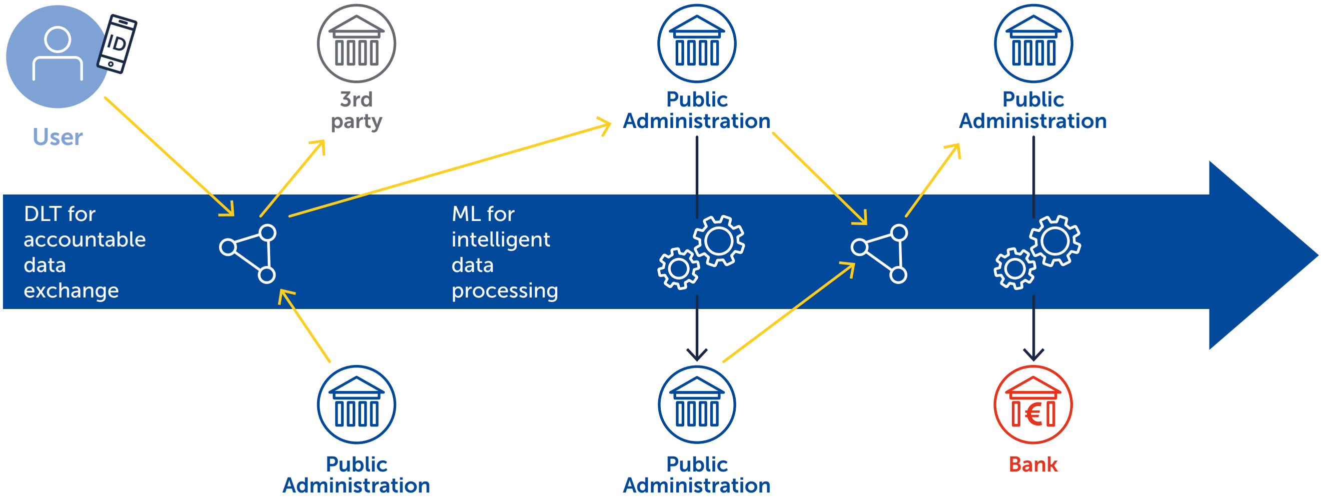


Figure 40.
Transforming public administration services into a proactive and non-interactive process

Benefits: Proactivity & non-interaction with DLT & AI

In the course of our research, we developed and piloted a readiness assessment method for public administration services that focuses on efficient processes, user-centered services and a system that appeals to public administration employees.

Main contributions of our project include:

- Concept: We provided a characteristics-based definition of proactive and interactionless government services, including business process and technology perspectives.

Analysis Framework: We developed a set of characteristics that should be studied along a structured method in order to the question: How to analyze government services towards AI-enabled proactivity and non-interaction?

- Demonstrator: We have demonstrated how our results can be applied to build a prototype for a particular government service.

Tangible assets of our work include:

- Publications (Balta et al, 2019; Kuhn et al, 2020a; Kuhn et al, 2020b)
- Talks & presentations with relevant stakeholders
  – Winner of the 2nd place for innovative concepts for government services at the National Science Dialog
- Demonstrator: Code and video (Demo, 2020)
- Evidentia: Enhancing the existing framework towards trusted decentralized cloud infrastructure for the government services of the future (Evidentia, 2020)
- An architectural perspective of government digitization in Germany and potential data interfaces that can be used to enhance services through AI (DigiGov, 2020)

Directions and challenges

We structure potential directions for future research based on existing challenges as outlined in the following table.

	Administration	Modeling	Processing	Communication & Interaction	Security & Privacy
Organiza-tional	What does a service provision process with ML look like.	What is a reference model or an architecture.	How to integrate AI into existing tools for the design of government services.	How to assure a technology shift.	How can GDPR conformity be evaluated and assured.
Semantic	What are shared context-specific concepts that can be integrated using ML.	What is a suitable ontology.	What are available ML tools & services that fit the requirements.	How to share language specific AI results.	How to exchange rules.
Technical / Syntactic	How to monitor and manage ML applications.	How to integrate AI applications into existing IT infrastructures.	How to improve data quality and customize models.	Which competencies are required for the application of ML.	What are suitable certification tools.

LITERATURE

Balta, D., Kuhn, P., Sellami, M., Kulus, D., Lieven, C., & Krcmar, H. (2019). How to streamline AI application in government? A case study on citizen participation in Germany. In International Conference on Electronic Government (pp. 233-247). Springer, Cham.

Demo (2020): Code for the backend https://git.fortiss.org/c4ai-drpm Video capture available at https://youtu.be/ZC3Smt54K1I

DigiGov (2020): https://digigov.fortiss.org/

Evidentia (2020): https://git.fortiss.org/evidentia

Kuhn, P., Balta, D., & Krcmar, H. (2020a). Was sind Herausforderungen proaktiver Verwaltungsleistungen in Deutschland. Wirtschaftsinformatik.

Kuhn, P., & Balta, D. (2020b). Service Quality Through Government Proactivity: The Concept of Non-interaction. In International Conference on Electronic Government (pp. 82–95). Springer, Cham.

3.2.3. Human-centered Machine Learning

Authors:
Dr. Yuanting Liu, Dr. habil Hao Shen, Sören Klingner, Zhiwei Han, Stefan Matthes, Tianming Qiu

Development of improved personalized stress detection models and a VR stress simulation for firefighters to create new datasets.

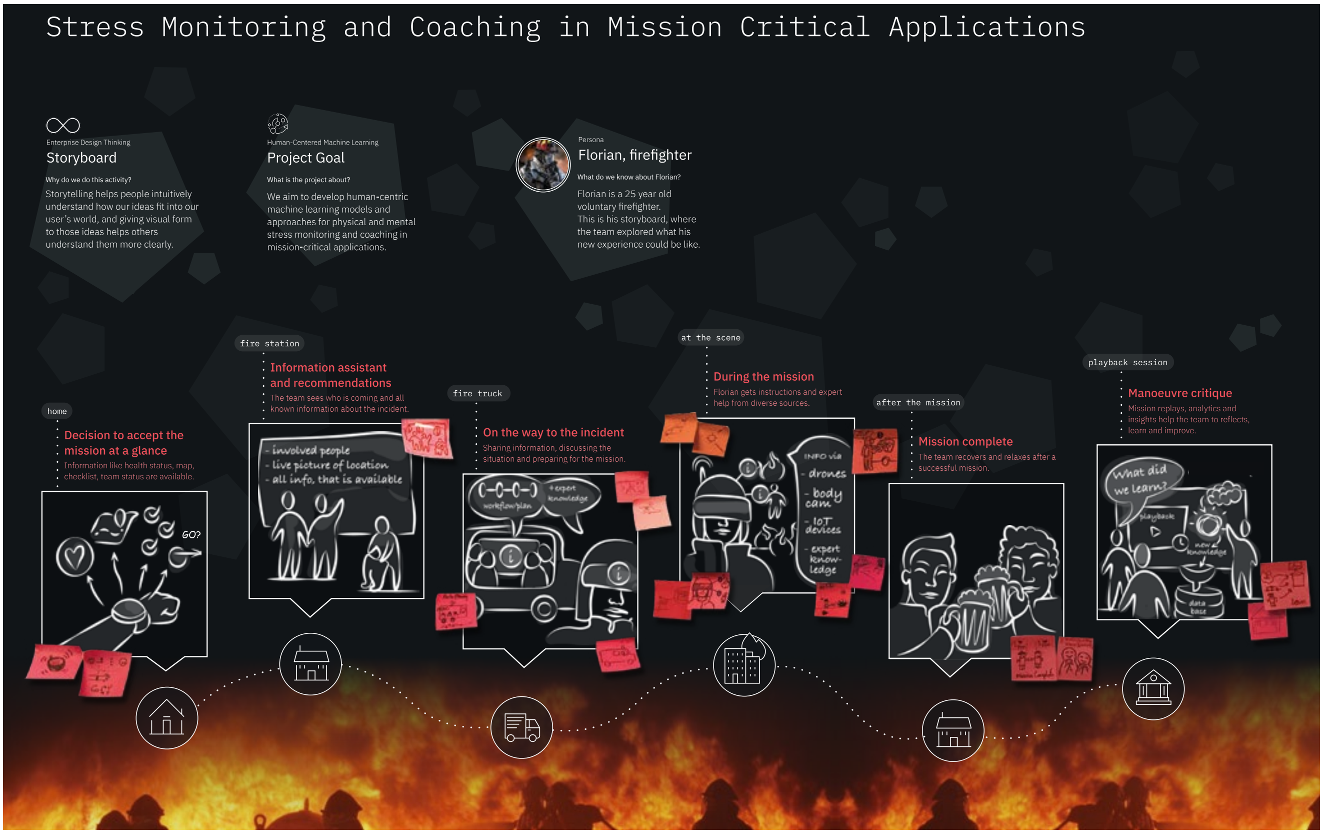
Firefighters work under immense pressure when responding to an emergency call. Stress up to a certain level increases alertness and productivity and can therefore be considered positive, while high stress usually leads to decreased productivity, impaired decision-making ability, reduced situational awareness, and life-threatening symptoms such as improper firefighting decisions. The accurate detection of intolerable stress thus contributes to reliable stress management, improved team performance and reduced risk to individuals during dangerous operations. Extreme heat, smoke that obstructs vision, time pressure and hazards are all factors under which firefighters must respond. The stress caused by such situations impairs responsiveness, both physically and mentally, and has potentially serious effects on cognitive abilities. Those affected are often unaware of their impaired judgment, which can have fatal consequences, as any mistake can cost their own or another person’s life. It is therefore desirable to develop a solution to reduce the risk caused by high stress levels.

To address this challenge, IBM and fortiss committed to working together to develop a modern data-driven “Stress Management for Firefighters” solution with a user-centered design approach and machine learning techniques at the Center for AI Research.

User-centered development as key

To understand the working environments, determine mission tasks and identify the needs of the firefighters, we organized a design thinking workshop with volunteer firefighters in Munich with the intent of addressing the mission-critical challenge in a user-centered way. This encouraged us to designate various scenarios, diverse critical tasks and potential solutions. We created a storyboard (see Figure 41) to convey the mission, which helped us intuitively understand when high stress levels can be risky, how mental or physical stress influences the situation and what type of support is expected. Furthermore, an archetypal user, utilizing a persona method, was created to represent the goals and needs of the firefighters, which enabled us to make informed decisions in the development of realistic scenarios from the start.

During the workshop discussions, we determined that detection accuracy is restricted by several factors, such as detection methods that are too generalized, the wearability of sensors and the real-time performance of detection algorithms. To address those challenges, an insightful and multi-disciplinary discuss session was initiated and a draft personalized stress detection solution for human-centered machine learning (HCML) was proposed as the most applicable approach. Meanwhile, one factor that has to be taken into account is the individual reaction to stress, which depends not only on the person’s current physical and mental



Stress Monitoring and Coaching in Mission Critical Applications

Enterprise Design Thinking
Storyboard
Why do we do this activity?
Storytelling helps people intuitively understand how our ideas fit into our user's world, and giving visual form to those ideas helps others understand them more clearly.

Human-Centered Machine Learning
Project Goal
What is the project about?
We aim to develop human-centric machine learning models and approaches for physical and mental stress monitoring and coaching in mission-critical applications.

Persona
Florian, firefighter
What do we know about Florian?
Florian is a 25 year old voluntary firefighter.
This is his storyboard, where the team explored what his new experience could be like.

home
Decision to accept the mission at a glance
Information like health status, map, checklist, team status are available.

fire station
Information assistant and recommendations
The team sees who is coming and all known information about the incident.

fire truck
On the way to the incident
Sharing information, discussing the situation and preparing for the mission.

at the scene
During the mission
Florian gets instructions and expert help from diverse sources.

after the mission
Mission complete
The team recovers and relaxes after a successful mission.

playback session
Manoeuvre critique
Mission replays, analytics and insights help the team to reflects, learn and improve.

Figure 41.
A storyboard for the firefighter application

capacity, but also on the situation while the measurement is carried out. In other words, a practical stress detection model should be able to identify inter- as well as intra-individual differences induced by different physical and psychological conditions such as gender, age, individual stress tolerance, and health status, which influence how humans experience stress.

By recognizing and combating firefighter stress, an efficient way to achieve better detection method generalizability is to personalize stress detection by capturing inter-individual differences. Stress detection personalization methods from prior research suffer from low personalization data efficiency which among other things is due to the huge data amount required for the identification of correlations between user-profiles and personalization models.[135]

Collecting data with virtual reality

To close this gap, fortiss investigated ways to measure and estimate the stress level of firefighters in real-time with the aim of assisting mission commanders in critical decision making. We addressed the result and knowledge gained from the design thinking workshop in developing accurate detection models. Our established studies with the "human in the loop" method showed significantly improved detection efficiency with only very few manual annotations (Weber et al, 2020). On this basis, fortiss conducted a virtual reality (VR) firefighting simulation by offering real, critical missions for multilevel stress detection models, which facilitate the trigger of person-specific physiological data and corresponding scene information in a controlled environment (Klingner et al, 2020). For the underlying stress data, we utilized various biosignal-sensors to record general stress indicators such as heart rate, brain activity, muscle tension and skin moisture as task inputs (as shown in Figure 42).

[135] Some research on stress detection personalization methods:

Sharma, Nandita, and Gedeon, Tom. "Stress classification for gender bias in reading." International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, 2011.

Shi, Yuan, et al. "Personalized stress detection from physiological measurements." International symposium on quality of life technology. 2010.

Schmidt, Philip et al. "Wearable affect and stress recognition: A review." arXiv preprint arXiv:1811.08854 (2018).

Nkurikiyeyezu, Kizito et al. "The influence of person-specific biometrics in improving generic stress predictive models." arXiv preprint arXiv:1910.01770(2019).

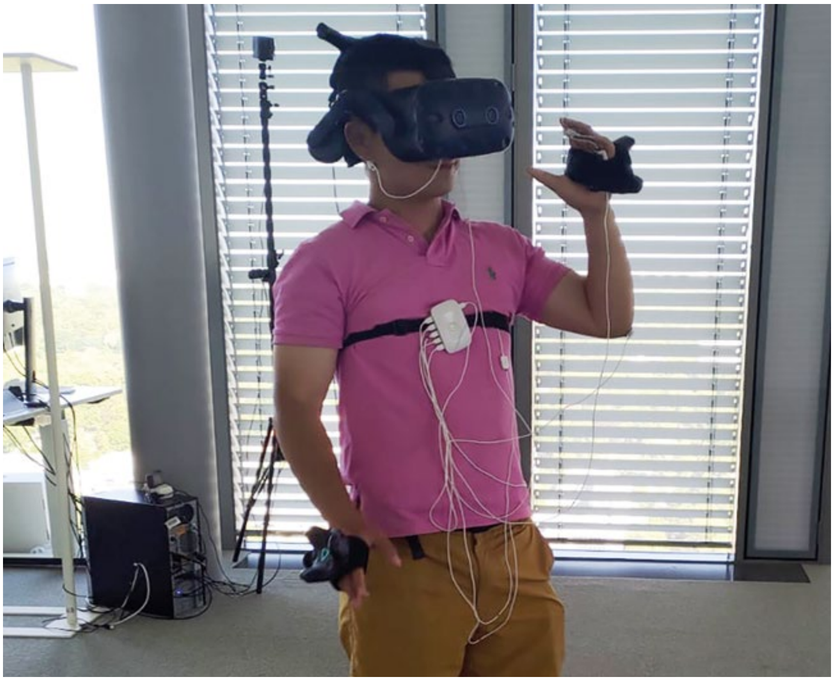


Figure 42.
VR setup with biosignal-sensors

This enables us to develop new stress recognition models by means of various firefighting scenarios and experience gained from simulated missions. With this simulation (experience), the user can physically walk through a burning apartment in a restricted tracking space without the need for a controller, which is achieved by using un-noticeable portals to redirect the user back into the center of the tracking space (see Figure 43). The multilevel stress setups can be configured and consist of different levels of mental stressors (visual, audio, time pressure, navigation) and physical stressors (walking, crouching, crawling, weights). This novel study was honored with the best poster award at the 43th German Conference on Artificial Intelligence in 2020.

Human-in-the-loop method with self-supervised learning

As part of our current results, our approach of personalization based on self-supervised learning technique (SSL) efficiently explores inter- and intra-individual differences with the "human in the loop" with the purpose of reducing the bias induced by the inaccurate labels. Most existing ML-based stress detection methods use a supervised classification approach and suffer from the poor label quality caused by non-standardized definitions of stress and varying stress resilience degrade detection performance. The main reason is that the data from the training set and the test subjects are not independent and identically distributed. Our approach consists of four sampling principles for manual labelling, which help to avoid the bias induced by unreliable labels in datasets. Subsequently, personalization is achieved by recalibration with active human feedback and efficient interaction. Compared to the conventional supervised-learning technique, our approach achieved significantly high accurate stress detection results (Matthes et al, 2020) through the development of a label-free SSL-based feature extractor.

Future activities and challenges

Our HCML algorithms for monitoring stress, based on data mining and cognitive characteristics, aim to deliver a clearly user-understandable stress state. The challenge with this type of data-driven research is always the quality and quantity of the data. We are searching for better ways to achieve personalized and standardized label acquisition for combined physical and mental stressors. By continuing to examine how to combine user modeling with interactive approaches, we will be in a position to determine the relationship between the user profile and inter-individual variability.

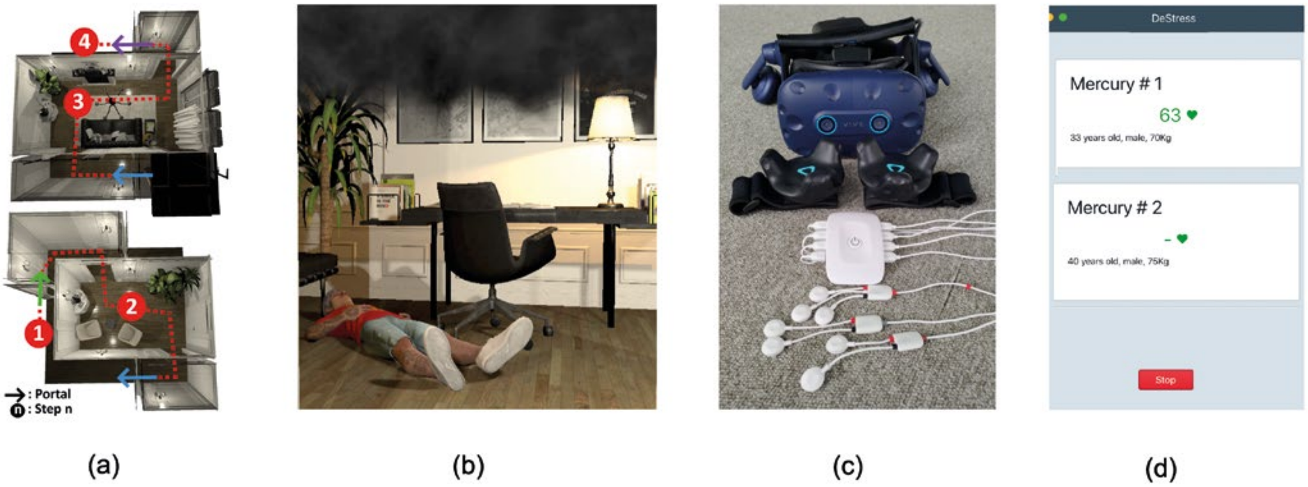


Figure 43. (a) Overview of the participants path inside vr; (b) Fire victim inside vr; (c) htc vive pro eye and biosignalplux sensorkit; (d) User interface of participants current stress level

As a framework of processes, a user-centered approach helps enhance the development of intelligent systems, particularly for mission-critical applications. The current research results will thus be reviewed together with professional firefighters in a second workshop. Their assessments wil help us to refine our research into personalized systems that identify at-risk firefighters and diminish stress and dysfunction. Our HCML approach and algorithms, which will be further validated in the field, extend the functionality of decision support systems for mission commanders and in other security-critical applications such as law enforcement.

LITERATURE

Klingner, S. , Han, Z. , Liu, Y. , Fan, F. , Altakrouri, B. , Michel, B. , Weiss, J. , Sridhar, A. , Chau, S. (2020). Firefighter Virtual Reality Simulation for Personalized Stress Detection. In German Conference on Artificial Intelligence (Künstliche Intelligenz) (pp. 343–347). Springer Verlag.

Matthes, S. , Han, Z., et al. (2020): Personalized Stress Detection with Self-supervised Learned Features. Human in the Loop Learning Workshop, 38th International Conference on Machine Learning. 2020.

Weber, T., Han, Z. , Matthes, S. , Liu, Y. , Hussmann, H. (2020): Draw with me: human-in-the-loop for image restoration. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. 2020. pp. 243–253.

3.2.4. Anomaly detection in robot-based manufacturing with semantic digital twins

Authors:
Dr. Markus Rickert, Alexander Perzylo

Industrial manufacturing is a very competitive market, in which the operational costs dominate the total cost of ownership of robot-based production systems. These costs include the efforts for setting up, reconfiguring, and handling the systems during operation. In order to reduce operational costs, while coping with the growing demand for customized products, the level of autonomy of manufacturing systems needs to be increased (Perzylo et al, 2019a). This includes not only the programming of the control logic of robots and their tools, but also the configuration of analytics mechanisms that are designed to monitor the execution of the production processes. No technical system is free of errors; hence an autonomous production environment has to be able to recognize and handle errors in an automated manner. Through the continuous monitoring of process parameters and sensor data from the involved manufacturing resources, anomalies can be automatically detected and analyzed. This enables the design and implementation of coping strategies, which help to increase the production system's resilience toward uncertainties and external influences.

Traditional robot-based automation requires an explicit specification of individual low-level commands in order to achieve a certain result. This type of programming by extensively trained experts is only commercially viable for industrial applications in mostly static high-volume settings that do not require frequent (re)programming. Under these circumstances, anomaly detection solutions can be manually tailored to a particular process and trained with a wealth of sensor data, as production phases tend to be rather long.

In contrast, small-scale production and the manufacturing of personalized goods result in a highly dynamic production environment that requires frequent changeover. The traditional programming approach is unsuitable given the changes in production requirements. It is not feasible to manually adapt manufacturing processes to handle a large quantity of product variants. In order to address these changes, a modern robot-based manufacturing system is required to autonomously factor in the production goals and capabilities with its own reasoning faculties. It must be able to automatically derive manufacturing plans and schedules and to assign individual tasks in a manufacturing process to compatible actors, such as devices or human workers. Rather than manually implementing program sequences for each involved resource, the specification of the product and its manufacturing process now form the driving aspect of the robot system. For making informed decisions, the system requires access to all relevant information such as the target product, its associated manufacturing process, and all available production resources.

Given these dynamic demands, anomaly detection approaches can no longer be manually prepared and optimized, as the exact order of activities is not known beforehand and is automatically generated based on the process and product specifications. Small and medium-sized enterprises (SMEs) in particular lack the expertise in technologies that are required for establishing anomaly detection solutions in their production environments. They cannot properly assess the type and quantity of data or the type of machine learning technique that is suitable for their production issues. Hiring external experts to take care of these issues is cost-intensive and not feasible for small production volumes. Small batch assemblies further pose the problem of data scarcity. Anomaly detection solutions must be initialized with data from only a few production cycles and should be continuously trained to improve over time. For supervised machine learning approaches, collected data must be labelled, in order to assign meaning to raw sensor values. In many scenarios, this is still a time-consuming manual process. And in contrast to traditional predictive maintenance applications, anomaly detection in small-volume and high-variety manufacturing must be able to cope with unstructured production environments. Instead of repeatedly performing the exact same motions in combination with fixed part locations, robots execute similar but not identical motions based on sensory input, such as from visual object detection.

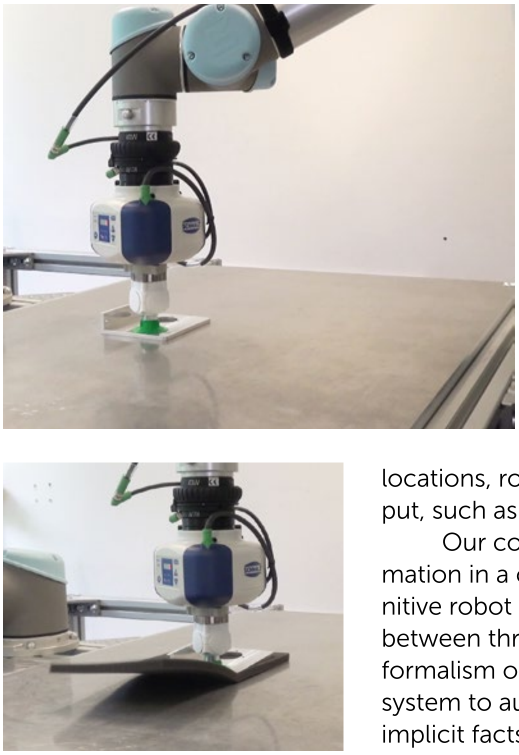


Figure 44.
Example: Nominal execution of a pick and place task (top), and a location anomaly regarding the interaction object due to a cluttered workspace (bottom)

Our concept aims to describe the relevant aspects of manufacturing automation in a common semantic description language that is interpreted by a cognitive robot system. It is based on the widely used PPR model that distinguishes between three major entity types: process, product, and resource. The underlying formalism of the semantic representation is a description logic that permits the system to automatically validate the logical consistency of models and derive implicit facts from explicitly modeled ones.

A key feature of this approach is to make manufacturing knowledge explicit that is often only available from employees, software implementations, or unstructured documents. By doing so, maintaining and reusing this knowledge can be more easily accomplished. Moreover, technical systems are enabled to automatically process the knowledge, in order to achieve a higher degree of autonomy and system resilience. For describing products and manufacturing resources alike, we rely on ontologies that describe geometric properties and inter relational constraints. Geometry models follow a BREP paradigm, in which the faces, edges, and vertices of an object are specified through exact mathematical models instead of polygon-based approximations. Utilizing such a semantically rich representation of all relevant entities leads to many synergy effects. For instance, process specifications can refer to the geometric properties of involved tools and products to semantically define process parameters. We call individual device or component models *semantic digital twins* that can be used as building blocks within larger work cells or factory models, in which device instances are placed and connected in a production environment (Perzylo et al, 2019b). Their poses as well as topological connections are semantically encoded and can be queried to analyze the flow of materials.

Based on the rich semantic context information provided by the semantic digital twins of all manufacturing resources in the production environment, sensor data that is generated during production runs is automatically annotated with relevant information. For instance, a measured force is put in relation to the involved robot, tool, target object, and task description for instance (Perzylo et al, 2020). As a result, the training of machine learning-based anomaly detectors can be provided with automatically labelled data and sophisticated context information. The trained anomaly detectors can then assess for each new sensor data sample, whether the current situation represents the nominal execution of the manufacturing process or must be considered an anomaly. The production system may decide to either handle the anomaly itself or to involve a human worker (Ba et al, 2020).

The knowledge-augmented anomaly detection concept was implemented in a real robot work cell at fortiss. The work cell is comprised of an industrial robot arm, a tool changer, and different grippers and other tools (Figure 44). It features the automatic discovery of devices and their functionalities and demonstrates the automated configuration of a robot-based assembly process and anomaly detection pipeline. It further shows how semantic digital twins can be employed in a plug-and-produce system to enable flexible small-batch manufacturing that complies with the needs of SMEs (Profanter et al, 2021). As a first use case, the manipulation of aluminum parts, such as through pick and place operations, was considered. Based on a limited amount of nominal execution runs, a corresponding model is trained, which is then used in subsequent runs to assess the state of the robot system. The system has been used to detect anomalies, such as part location anomalies, obstacles on the robot path, or vacuum leakages for the vacuum gripper. In the next project phase, we are aiming to establish collaborations with manufacturing SMEs for testing our concepts in additional test beds (Rickert and Ploennigs, 2020).

LITERATURE

Ba, A., Profanter, S., Perzylo, A., Altakrouri, B., Rickert, M. , Ploennigs, J. (2020) “A neuro-symbolic AI approach to anomaly diagnosis for robot-based manufacturing settings”, Video Publication, 2020, https://youtu.be/cgx8NOT_c1M.

Perzylo, A., Kessler, I., Profanter, S., & Rickert, M. (2020). Toward a Knowledge-Based Data Backbone for Seamless Digital Engineering in Smart Factories. In 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (Vol. 1, pp. 164–171). IEEE.

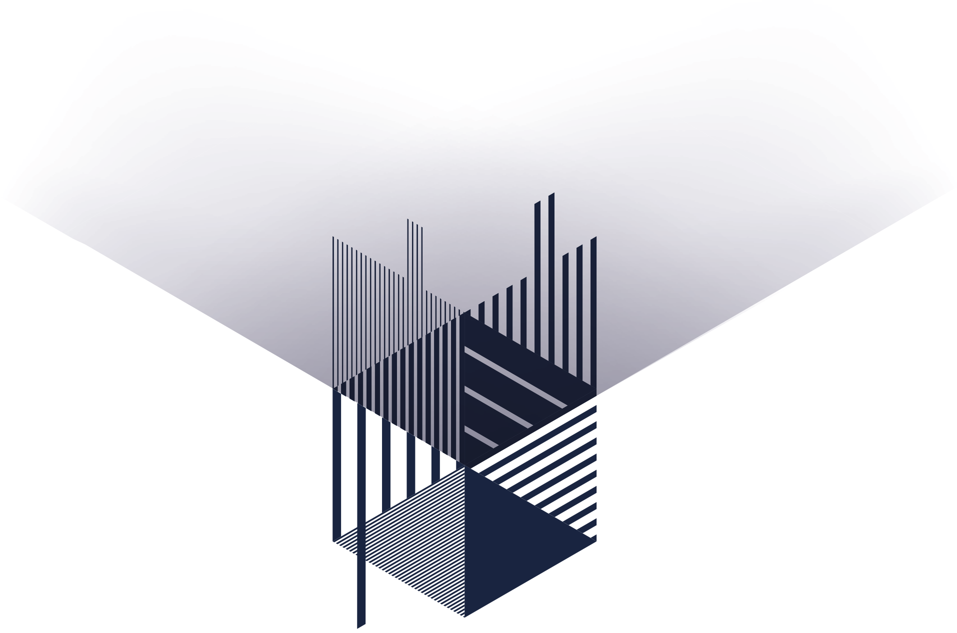
Perzylo, A., Rickert, M., Kahl, B., Somani, N., Lehmann, C., Kuss, A., ... & Danzer, M. (2019a). SMErobotics: Smart robots for flexible manufacturing. IEEE Robotics & Automation Magazine, 26(1), 78–90.

Perzylo, A., Profanter, S., Rickert, M., & Knoll, A. (2019b). OPC UA nodeset ontologies as a pillar of representing semantic digital twins of manufacturing resources. In 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) (pp. 1085–1092). IEEE.

Profanter, S., Perzylo, A., Rickert, M., & Knoll, A. (2021). A Generic Plug & Produce System Composed of Semantic OPC UA Skills. IEEE Open Journal of the Industrial Electronics Society.

Rickert, M. , Ploennigs, J. (2020): Roboter für die Kleinserie: Digitale Zwillinge eröffnen Möglichkeiten für den Mittelstand”, IBM THINK Blog DACH, https://www.ibm.com/de-de/blogs/think/2020/12/08/digitale-zwillinge-fuer-roboter/.

» We are able to show interested visitors, researchers, application partners and networks how we can shape future developments and exploit the potential associated with digitization. «



4

TRANSFER

TRANSFER

4.1 fortiss Mittelstand

Author:
Dr. Wolfgang Köhler

The mission of fortiss Mittelstand is to prepare SMEs for the future in the area of software and AI with a comprehensive range of information, qualification and implementation services. The service portfolio offers needs-based support for all questions concerning AI, including for companies that are at the beginning of the digital transformation process.

At the location on the 15th floor of the Highlight Tower in Munich and in co-operation with AI-relevant competence fields, SMEs receive support in potential analysis, prototyping and validation of new AI-based products and services. This offer is supplemented by the installation of demonstrators from various fields of competence in the fortiss lab in the southern part of the 15th floor.

The range of services offered by fortiss Mittelstand is divided into "Informing", "Qualifying" and "Implementing". This division into three parts roughly reflects the different needs of SMEs, depending on how intensively they have already dealt with the topic of software and AI.

In the "Informing" section, decision-makers from SMEs are introduced to current trends and technologies in the field of software and AI in one-day networking events and conferences. Within the framework of these events, participants learn through interactive workshops and the presentation of real-life applications how the potential of software and AI can be used.

In the area of "Qualification", fortiss Mittelstand organizes further education programs that focus on a research topic that is considered relevant for SMEs. In terms of the educational content, these events are led and supported by the corresponding research fields from fortiss.

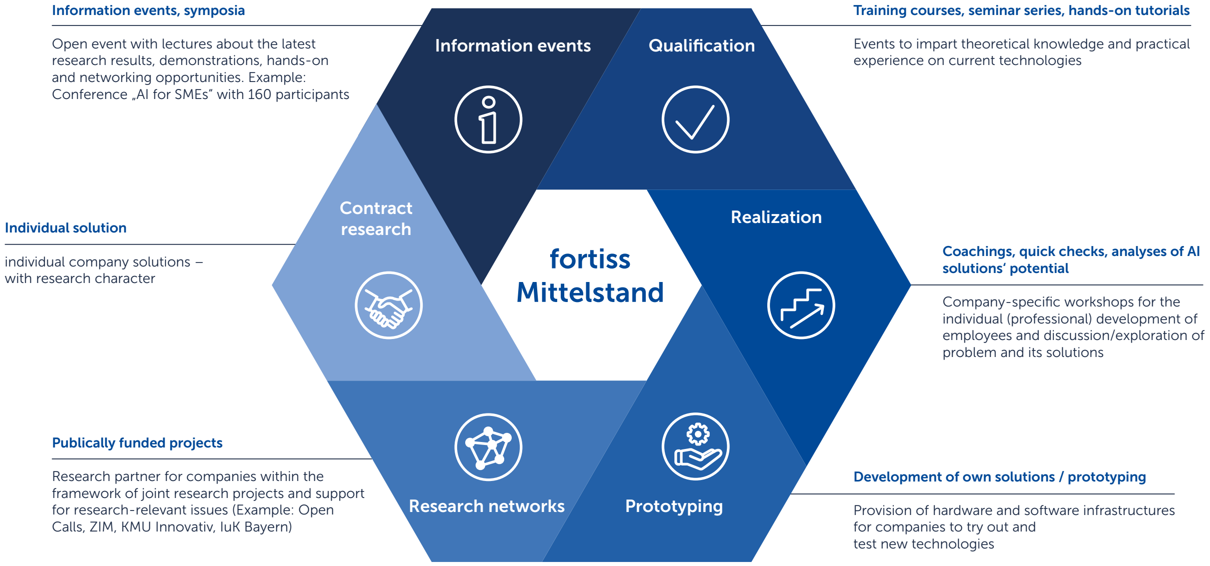
The "Implementation" area refers to customized solutions for individual companies. fortiss Mittelstand offers company-specific workshops for the individual development of employees and the discussion or exploration of problems and solution options. Furthermore, fortiss provides hardware and software infrastructures for prototyping workshops to try out and test new technologies. fortiss is also a research partner for funded AI projects. The Free State of Bavaria and the Federal Republic of Germany offer competitive funding opportunities with rapid evaluation, especially for SMEs. Examples include: ZIM[136], KMU-innovativ[137] IuK Bayern[138].

As part of this tripartite service portfolio, the following webinars and webinar series were successfully conducted:

[136] www.zim.de

[137] www.bmbf.de/de/kmu-innovativ-561.html

[138] www.iuk-bayern.de



Information events, symposia

Open event with lectures about the latest research results, demonstrations, hands-on and networking opportunities. Example: Conference „AI for SMEs" with 160 participants

Training courses, seminar series, hands-on tutorials

Events to impart theoretical knowledge and practical experience on current technologies

Individual solution

individual company solutions – with research character

Coachings, quick checks, analyses of AI solutions' potential

Company-specific workshops for the individual (professional) development of employees and discussion/exploration of problem and its solutions

Publically funded projects

Research partner for companies within the framework of joint research projects and support for research-relevant issues (Example: Open Calls, ZIM, KMU Innovativ, IuK Bayern)

Development of own solutions / prototyping

Provision of hardware and software infrastructures for companies to try out and test new technologies

Figure 45.
Service portfolio for SMEs

- In cooperation with the Lars and Christian Engel Foundation (LUCE), Weiherhammer, Oberpfalz, the online webinar "Best-practice digitization and AI for SMEs" took place on 19 November 2020. The Denkwelt Oberpfalz and the Überbetriebliche Bildungszentrum in Ostbayern (ÜBZO) were presented, as well as funding opportunities for AI cooperation projects between companies and universities / research institutions by the project executing organisation VDI/VDE. BAM GmbH also provided a best practice presentation on the topic of AI implementation in the company.
- In cooperation with the Chamber of Commerce, Munich, and UnternehmerTUM, a five-day webinar series on artificial intelligence took place in October and November 2020. The aim of this webinar series was to illustrate examples of how SMEs in particular can get started with the topic of artificial intelligence. The benefits and potential of the technology for companies were explained and the development of an AI strategy was shown. For this purpose, basic technical content was combined with strategic perspectives and recommendations for action for the implementation of AI activities in one's own company.
- For members of the VDMA Bavaria association, another webinar series took place in December 2020 on the topic of "Deep Dive Artificial Intelligence". The aim was to give participants a quick but targeted overview of AI and related disciplines. This webinar series provided participants with an intuitive approach to handling and using data. Accordingly, they were able to independently identify initial potential within the company and prototype them through a "proof of concept". The idea behind this fortiss webinar concept is to divide the learning content into three superordinate dives, where a specific topic is addressed in detail. The different dives—promoting intuition in dealing with data and models, technical implementation and prototyping, identification of suitable use cases in SMEs - build on each other and involve the participants actively through a corresponding online "hands-on" component.

The fortiss Mittelstand service portfolio is also part of the DIH Munich Innovation Hub for Applied AI, which fortiss operates together with UnternehmerTUM and MSRM. This DIH is part of the DIHNET.EU project, an association of numerous DIHs across Europe that build a "network of networks" around digital transformation initiatives. It is also part of the AI DIH-Network, so listed in the European Catalogue with technological focus on AI and on other technologies that demonstrate interaction with AI. In this way, fortiss Mittelstand opens up the activities to all SMEs across Europe. This makes it possible for Bavarian SMEs to network with research institutes and industry partners across Europe. Via special "Open Call" cascade funding initiatives, the project consortium receives additional funding reserved exclusively for "Selected Third Parties", such as SMEs. The consortium defines topics, organizes calls for funding—open calls—, organizes independent reviews and supports winners in project implementation. With fortiss as a partner, a number of DIH-related research projects have already been launched, many of which offer transfer activities disseminated through DIH Munich. These include the following projects:

- HumaneAI: The aim of this EU project is to develop AI technologies that are beneficial to people and society and in line with Europe's ethical values and social, cultural, legal and political norms.
- DIH4AI: The DIH4AI project aims to build a network of AI-on-demand innovation and collaboration platforms that will ensure the co-development and delivery of ecosystem, business, technology and transformation services through a sustainable network of regional DIHs specialized in AI across Europe. fortiss plays a significant role here in providing services and experiments on the topics of edge ecosystem services, "HumaneAI" innovation ecosystem, Smart Energy Living Lab, out-of-the-box platform-as-a-service for accountable evidential transactions and an online AI evaluation check. These services and experiments will be realized with selected European SMEs - identified and supported in the DIH4AI project through two cascade funding processes.
- HUBCAP: HUBCAP is a so-called innovation action within the European Commission's "Smart Anything Everywhere" initiative. The goal of Smart Anything Everywhere (SAE) is to help SMEs, start-ups and mid-caps enhance their products and services through the inclusion of innovative digital technologies.
- VOJEXT: Under the vision Value Of Joint EXperimentation (VOJEXT) in digital technologies, this project dynamizes science-driven industry approaches engaging human and cyber-physical systems (CPS) in the same loop, thus amplifying the cognitive capabilities needed to achieve more effective sociotechnical and business ecosystems. For this purpose, VOJEXT will design, develop, and demonstrate affordable, market-oriented, multipurpose and easy-to-repurpose robotic systems.

So far, a total of around 60 SMEs have been supported by funding programs in which fortiss was directly or indirectly involved. Apart from our DIH-activities, fortiss Mittelstand has been able to expand its network worldwide. In September 2020, fortiss Mittelstand initiated a cooperation with the UC Berkeley School of Information that involved an online AI strategy course for SMEs.

4.2
Center for Code Excellence

Author:
Dr. Johannes Kross

The Center for Code Excellence (CCE) combines fortiss' expertise in the analysis, development and transfer of modern software engineering methods, tools and processes. It represents a contact point for Bavarian medium-sized companies and provides knowledge and services to develop outstanding, sustainable and future-oriented software. In doing so, CCE targets software developers as well as managers. Since the share of ML components in (traditional) software systems steadily increases, there is an opportunity and need to adapt and apply well-known software engineering methods to the development and integration of ML components. As a result, the CCE conducts research into open issues such as 1) techniques required to track and share data, source code, models and results 2) how to automate and orchestrate ML systems and ML system delivery 3) life cycles for organizing the testing, development and deployment of ML systems for different roles.

CCE offers this knowledge and research portfolio to outstanding master's and PhD students and, respectively, prospective entrepreneurs and newly-created start-ups within the scope of the TUM Venture Labs where CCE and fortiss are an integral part of the key TUM Venture Lab for Software and AI (SW/AI). The TUM Venture Labs are an initiative by TUM and UnternehmerTUM and represent several innovation hubs in the domains of engineering, natural, life and data sciences and medicine. They aim to foster and help young entrepreneurs in creating technology start-ups and business translation from research. As a result, they provide an entire ecosystem for development, training and network ventures. Since the mission of TUM Venture Labs is to incentivize technology-based start-ups, the SW/AI Venture Lab is a key platform. The mission of the SW/AI Venture Lab is to push scalable business ideas in the software, data and AI files, and build up a cross-functional SW platform to target interdisciplinary fields of the future. It offers various entrepreneur modules, event networks, spaces and infrastructures (data access, computing resources), and, lastly, educational programs.

fortiss offers support in educational programs targeted at the specific needs of software, AI or software-enabled start-ups, team matching and funding access. fortiss has successfully started the first educational offering of the SW/AI Venture Lab and launched a ML Training Camp for 20 master's and PhD students from different domains such as chemistry and sports and health. The course provides the mathematical foundation and basic technologies of machine learning, beginning with common tools for dimensionality reduction, to current deep learning approaches and reinforcement learning. The learning objectives are designed to answer questions such as "What is ML? How does it work? How do I benefit from ML?" and to develop and apply simple ML models. In addition, the SW/AI Venture Lab offers the aiSpace to auspicious entrepreneur teams, which collaborates with fortiss and UnternehmerTUM appliedAI in order to launch new ventures.

» We are compiling best practices on AI engineering, and contribute to corresponding standardization efforts. fortiss transfer centers offer a portfolio of information, qualification, and prototyping formats based on state-of-the-art findings and experience on AI Engineering. «

IMPRINT

Publisher

fortiss
www.fortiss.org
© 2021

Editors

Dr. Harald Rueß
Dr. Xin Ye

Layout

Sonja Taut

Proofreading

Daniel Hawpe

Print

viaprinto | CEWE Stiftung & Co. KGaA
Martin-Luther-King-Weg 30a
48155 Münster

Date of Issue

February 6th, 2021

Picture Credits

Title Illustration © Sonja Taut
P. 77 © fortiss
P. 79 Figure 23: © fortiss; Figure 24: © Astrid Eckert; Figure 25: © fortiss
P. 80 © fortiss
P. 82 © fortiss
P. 83 Figure 29: © Sonja Taut
P. 84 © fortiss
P. 86 © Sonja Taut
P. 88 © fortiss
P. 104 © fortiss
P. 107 © fortiss

fortiss is the Free State of Bavaria research institute for software-intensive systems based in Munich. The institute's scientists work on research, development and transfer projects together with universities and technology companies in Bavaria and other parts of Germany, as well as across Europe. The research activities focus on state-of-the-art methods, techniques and tools used in software development and systems & service engineering and their application with cognitive cyber-physical systems such as the Internet of Things (IoT).

fortiss is legally structured as a non-profit limited liability company (GmbH). The shareholders are the Free State of Bavaria (majority shareholder) and the Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.

Although this report was prepared with the utmost care and diligence, inaccuracies cannot be excluded. No guarantee is provided, and no legal responsibility or liability is assumed for any damages resulting from erroneous information.

fortiss GmbH
Guerickestr. 25
80805 Munich
Germany
www.fortiss.org
Tel.: +49 89 3603522 0
E-Mail: info@fortiss.org

Sponsored by

Bavarian Ministry of Economic Affairs,
Regional Development and Energy