ercim-news.ercim.eu Number 134 July 2023

ERCIM NEWS

Special theme:

eXplainable AI

Also in this issue

Research and Innovation:

Fibre-Optic Sensing for Road-Traffic Monitoring in Remote Areas

Editorial Information

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 2,000 printed copies and is also available online ,at https://ercim-news@ercim.eu.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
+33 4 9238 5010, contact@ercim.eu
Director: Dominique Hazaël-Massieux, ISSN 0926-4981

Contributions
Contributions should be submitted to the local editor of your country

Advertising
For current advertising rates and conditions, see https://ercim-news.ercim.eu/ or contact peter.kunz@ercim.eu

ERCIM News online edition: *https://ercim-news.ercim.eu/*

Next issue:
October 2023: Climate-Resilient Society

Subscription
Subscribe to ERCIM News by sending an email to en-subscriptions@ercim.eu

Cover illustration by jambulboy on pixabay.

JOINT ERCIM ACTIONS

SPECIAL THEME

Diversity Matters

by Léon Ouwerkerk, CWI, interviewed by Monica Divitini, as part of the ERCIM HR initiative to create awareness for diversity in the broadest sense.

Diversity remains an important topic. ERCIM News' Monica Divitini interviewed Léon Ouwerkerk of CWI. Ouwerkerk: "At CWI, we think it is important that everyone feels safe to be themselves in the workplace. That is one of the reasons for us to take actions to reach more diversity, equity and inclusion. I am glad to be asked to tell you more about our policies".

What role do you play in your organisation?

Actually I have two roles. As manager of the HR department of Centrum Wiskunde & Informatica (CWI), the national research institute for mathematics and computer science in the Netherlands, I have a responsibility for our HR policies including stimulating equity, diversity and inclusion. This concerns cultural, ethnic and/or religious background, gender, sexual orientation, capability for work and health, and age. At CWI, we think it is very important that everyone feels safe at work.

The other role is as LGBTIQ+ coordinator for our mother organisation the Dutch Research Council (NWO for short), which is more informal and personal, but also much appreciated in the organisation. LGBTIQ+ stands for homosexual, lesbian, bisexual, transgender, intersex and queer persons.

Maybe the latter already sounds heavy for some people, but actually we like to keep it light, positive and inclusive for everyone. For example, on several Coming Out Days we had cheerful rainbow cakes for everyone at CWI. A small gesture like that already shows that you give attention to this group. And during our Pride events, it is very important that straight allies are invited too and that we have diversity in speakers as well. We sometimes discuss difficult issues there, but the general tone should give energy.

To help us with LGBTIQ+ inclusion, NWO is a member of Workplace Pride, a not-for-profit foundation dedicated to improving the lives of LGBTIQ+ people in workplaces worldwide. Through Workplace Pride, we learn to shape our policies and gain knowledge from other members. We can join their activities as well.

Last year, we signed their Declaration of Amsterdam, showing our commitment in fostering a more inclusive workplace for our LGBTIQ+ employees. As part of this, it is important that the employer identifies and supports leaders and decision makers (including straight) that actively strive to create LGBTIQ+ inclusive working environments. But LGBTIQ+ employees themselves also have a responsibility to endeavour being visible at work and collaborate with their employers on diversity and inclusion, leading the way for all employees.

Being part of the community myself and at the same time "employer", I feel a double responsibility to provide a safe, comfortable, equal opportunity workplace and promote authenticity for LGBTIQ+ employees. Especially as a member of the management, it is important to be "out", in my case open about being gay, and actively strive to make things better.

Why do you think it is important to promote inclusion and diversity in research institutes and universities?

It is scientifically proven that diversity in organisations brings inspiration, creativity and innovation. But it is also important that academic organisations are in connection with and a representation of society, to be able to give answers to today's questions, which is not possible when you are a white, straight male stronghold.

On top of that, we have an interest in this as academic organisations. There is a war on talent going on. The labour market is tight and talent is scarce. Diversity is very important to leave no talent unused, but it also makes you more attractive as an employer, especially for the younger generation. There really is a win-win here!

Can you briefly explain some initiatives that your organisation has started to promote diversity and inclusion? Any initiative that you are particularly proud of?

Most of our LGBTIQ+ initiatives are focused on inclusion. At NWO, we organise Pride events where mostly LGBTIQ+ scientists, from mathematicians to astronomers, tell about their research and their personal life. This offers positive role models, network opportunities, and shows that the employer supports this (also with budget). Seeing the rainbow PR posters for these events is already supportive for the community.

I have also been lobbying for some time to make our centralised terms of employment more inclusive like including transition leave for transgender people,

special attention for rainbow families etc. I am proud of my colleagues that they did more than that: with the support of the unions as well, our entire terms of employment are now being carefully reviewed on all diversity issues in general.

We also made a diversity plan, with very diverse action points. Among others, at CWI, we are making our recruitment and selection process more open to diversity by offering a special toolkit and a bias training to management. To improve social safety, we are introducing bystander trainings for everyone, and we recently had a big event with actors playing socially-non-safe-scenes and a discussion panel on side of it.

Have you faced any challenge in promoting inclusion and diversity?
Surprisingly little when it comes to LGBTIQ+, but more in general we sometimes get asked if we are not exaggerating a bit. My opinion is that when the balance is so far off, you sometimes have to take some extra steps to correct the situation.

Are there lessons learned or best practices that you would like to share with other organisations that want to work around these themes?
You can't expect people to work on diversity on top of their regular work or in their spare time. If you think it is important, give some dedicated persons the hours to work on this. It shouldn't be volunteers' work.

One last tip: share your pronouns under your email. It implicitly shows that you support your LGBTIQ+ colleagues, and that you are a safe person for them.

Léon Ouwerkerk (he/him)
HR manager
CWI, The Netherlands

ERCIM Appoints Dominique Hazaël-Massieux as Manager of ERCIM EEIG

The ERCIM EEIG board of Directors is pleased to announce the appointment of Dominique Hazael-Massieux as the new manager of ERCIM EEIG, effective from June 2023. Dominique will be taking over from Philipp Hoschka, who has served as the manager since December 2016.

Over the years, Dominique has been involved in developing and running a number of W3C technical programs. He took a leading role in the Mobile Web Initiative, which aims to enhance the usability of the Web on mobile devices. He has also contributed to the advancement of Web Real-Time Communications, the fundamental technology enabling modern video conferencing. Furthermore, Dominique has spearheaded the work on the Immersive Web, a groundbreaking initiative that brings Virtual and Augmented Reality experiences to the Web platform. Additionally, he launched and has managed the W3C Community Management program, which strives to boost developer engagement and adoption of W3C standards. Notably, Dominique initiated the efforts required to make Web browsers capable of running Machine Learning models. Many of these significant endeavors have received support from research funding provided by the European Union.

In October 2022, ERCIM appointed Dominique on the newly formed W3C Board of Directors to bring his experience to ensure a smooth transition into W3C's new legal entity status. Dominique holds an engineering degree from Ecole Centrale Paris.

In expressing his gratitude, Bruno Sportisse, President of ERCIM EEIG and CEO of Inria, extended a warm appreciation to Philipp Hoschka for his commitment and significant contributions during his mandates to ERCIM and the overall progress of W3C in recent years.

The appointment of Dominique Hazael-Massieux as the new ERCIM manager marks a milestone in advancing the mission and objectives of ERCIM EEIG. Dominique's extensive experience and expertise within the W3C community make him an ideal candidate to lead ERCIM in its future endeavors.

Workshop on Digital Ethics in Research – Save the Date!

After the remarkable success of the Forum on Digital Ethics in Research held in October 2022, both in Paris and online, the ERCIM Ethics Working Group announces the upcoming edition of the workshop. Mark your calendars for the exciting event taking place on 18-19 October 2023 in Porto, held in conjunction with the ERCIM fall meetings.

For those who missed out or want to revisit the insightful presentations and slides from last year's workshop, the recorded sessions and slides are still accessible on the Beyond Compliance website. Be sure to visit the website, where we will soon provide more details about the 2023 workshop.

https://www.ercim.eu/beyond-compliance/

11th International Workshop on Computational Intelligence for Multimedia Understanding

by Behçet Uğur Töreyin (İTÜ), Maria Trocan (ISEP) and Davide Moroni (CNR-ISTI)

Around fifty researchers attended the International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), organized annually by the working group Multimedia Understanding through Semantics, Computation and Learning (MUSCLE) of the European Research Consortium for Informatics and Mathematics (ERCIM), which took place as a satellite workshop of IEEE ICASSP 2023 held in Rhodes, Greece, June 5, 2023. This year's workshop was also a huge success as previous venues with two oral sessions and one poster session, hosting eighteen papers authored by seventy-three researchers affiliated with nineteen institutions coming from twelve countries.

Multimedia understanding is an integral part of many intelligent applications in our social life, whether in our households or in commercial, industrial, service, and scientific environments. Analyzing raw data to provide them with semantics is essential to exploit their full potential and help us manage our everyday tasks. Nowadays, raw data typically come from a host of different sensors and other sources and are different in nature, format, reliability and information content. Multimodal and cross-modal analysis are the only ways to use them at their best. Besides data analysis, this problem is also relevant to data description intended to help storage and mining. Interoperability and exchangeability of heterogeneous and distributed data is a need for any practical application. Semantics is information at the highest level, and inferring it from raw data (that is, from information at the lowest level) entails exploiting both data and prior knowledge to extract structure and meaning. Computation, machine learning, statistical and Bayesian methods are tools to achieve this goal at various levels. The MUSCLE working group through IWCIM aims to address these emergent topics by growing a community of scientists and practitioners from the academy and industry. The mission is still ongoing. In particular, the twelfth IWCIM is anticipated to be organised in conjunction with IEEE - ISCAS 2024, to be held in Singapore on May 19-22, 2024.

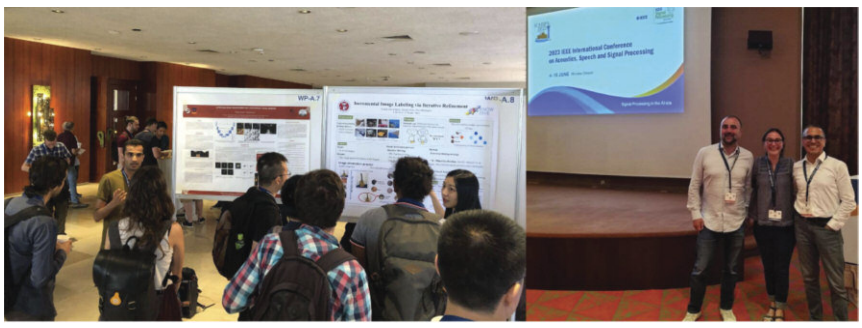
The IWCIM 2023 website is hosted by Istanbul Technical University, and full-text papers may be accessed via IEEExplore.

Links:
http://wiki.ercim.eu/wg/MUSCLE/
http://iwcim.itu.edu.tr

Please contact:
Behçet Uğur Töreyin, ITU, Istanbul,Turkey
toreyin@itu.edu.tr

Maria Trocan, Institut Supérieur d'Électronique de Paris (ISEP), Paris, France
maria.trocan@isep.fr

Davide Moroni, CNR-ISTI, Pisa, Italy
davide.moroni@isti.cnr.it



Highlights from the poster session and the workshop chairs.

ERCIM
European Research Consortium for Informatics and Mathematics

Horizon Europe Project Management

A European project can be a richly rewarding tool for pushing your research or innovation activities to the state-of-the-art and beyond. Through ERCIM, our member institutes have participated in more than 100 projects funded by the European Commission in the ICT domain, by carrying out joint research activities while the ERCIM Office successfully manages the complexity of the project administration, finances and outreach.

Horizon Europe:How can you get involved?
The ERCIM Office has recognized expertise in a full range of services, including:

- Identification of funding opportunities
- Recruitment of project partners (within ERCIM and through ournetworks)
- Proposal writing and project negotiation
- Contractual and consortium management
- Communications and systems support
- Organization of attractive events, from team meetings to large-scale workshops and conferences
- Support for the dissemination of results.

Please contact:
Peter Kunz, ERCIM Office
peter.kunz@ercim.eu

Workshop on Privacy, Transparency, Sovereignty and Security: A Summary

by Peter Kunz (ERCIM Office)

A collaborative workshop on privacy, transparency, sovereignty, and security was held on 27-28 April 2023 at the Inria Mediterranean Campus in Sophia Antipolis. The workshop was co-organized by the European projects TRAPEZE [L1] and E-CORRIDOR [L2], and co-sponsored by Inria and ERCIM. The two-day workshop aimed to explore new ideas, share insights and collaborate on solutions that can benefit society in the fields of privacy, cybersecurity, and technology. This report presents a brief overview of the presentations and discussions held during the workshop [L3].

The first day of the workshop was dedicated to research and technology. Alexander Vasylchenko from TenForce started the presentations by discussing how TRAPEZE services can enable digital privacy, security, and identity. This was followed by Fabio Martinelli's presentation on the E-CORRIDOR approach for confidential analytics. The next presentation was by Sabrina Kirrane from the Vienna University of Economics and Business, who discussed blockchain-based resource governance for decentralized web environments. Jean-Paul Bultel from the French Alternative Energies and Atomic Energy Commission-CEA presented homomorphic encryption-based similarity matching for privacy-preserving interest-based data sharing.

Ilaria Matteucci from the Italian National Research Council discussed privacy issues and the automotive sector. Simone Fischer-Hübner from Karlstad University presented a talk on usable transparency and consent, which was delivered by videoconferencing. Roland Rieke from Fraunhofer-Gesellschaft presented machine-learning methods for in-vehicle intrusion detection, with access restricted to the presentation. João P. Vilela from INESC TEC, CISUC & University of Porto talked about the prediction of user privacy preferences in mobile devices via federated learning. The last presentation of the day was by Francesco di Cerbo and Volkmar Lotz from SAP, who jointly presented on privacy research topics at SAP. The day ended with lively discussions during a wine and cheese reception.

The second day of the workshop focused on two main areas: privacy policy and regulations and the innovative utilization of privacy technologies across different sectors. The day started with a two-hour session that brought together research and Data Protection Authorities (DPAs) as well as experts from the data protection community to foster synergies and collaboration. The discussion explored the role of DPAs in promoting data protection and privacy across different sectors and jurisdictions. The session also discussed selected privacy-enhancing results of the TRAPEZE project, such as consent management through a citizen-centric privacy dashboard, policy language, and the concept of sticky consent policies.

After the DPA session, presentations were given on the challenges of trust and sovereignty awareness in global transactions, privacy and legal issues in big data, privacy-preserving passenger processing, data usage control, and protecting data and privacy in the public transport sector.

Martin Kurze (Deutsche Telekom), Ramon Martín de Pozuelo (CaixaBank), and Lauro Vanderborght (Digital Flanders) gave an overview of the TRAPEZE use cases & their impact on society. Theo Dimitrakos from Huawei and University of Kent) gave a presentation titled "Challenges of trust and sovereignty awareness for global transactions in an increasingly fragmented polarized world" via video conferencing. Rigo Wenning (ERCIM/W3C) talked about "Privacy and legal issues in Big Data." Stefano Sebastio, from Collins Aerospace presented privacy-preserving passenger processing and operations solutions for multi-modal travels, followed by Paolo Mori (Italian National Research Council) who introduced a Data usage control procedure via video conferencing. Fabio Podda (Azienda Mobilità e Trasporti Genova) and Liivar Luts (Tallinna Transpordiamet) jointly talked about how protecting data and privacy in public transport sector in the frame of the CitySCAPE project.

Alexander Vasylchenko and Fabio Martinelli, the coordinators of the TRAPEZE and E-CORRIDOR projects respectively, concluded the workshop summarising the topics and challenges presented and questions moving forward.

More than 70 participants attended the workshop, half of them online.

Workshop participants. Photo: P. Kunz.

Links:
[L1] https://trapeze-project.eu/
[L2] https://e-corridor.eu/
[L3] https://kwz.me/hxd

Please contact:
Peter Kunz, ERCIM Office
peter.kunz@ercim.eu

Introduction to the Special Theme

Explainable AI

by Manjunatha Veerappa (Fraunhofer IOSB) and Salvo Rinzivillo (CNR-ISTI)

Artificial Intelligence (AI) has witnessed remarkable advancements in recent years, transforming various domains and enabling groundbreaking capabilities. However, the increasing complexity of AI models, such as convolutional neural networks (CNNs) and deep learning architectures, has raised concerns regarding their interpretability and explainability. As AI systems become integral to critical decision-making processes, it becomes essential to understand and trust the reasoning behind their outcomes. This need has given rise to the field of explainable AI (XAI), which focuses on developing methods and frameworks to enhance the interpretability and transparency of AI models, bridging the gap between accuracy and explainability.

Explainable AI Methodology

The lack of transparency in AI models can hinder their effectiveness and introduce potential vulnerabilities. XAI aims to address this challenge by incorporating interpretability techniques into AI models, allowing security analysts and stakeholders to understand the reasoning behind AI-driven decisions. Héder discusses the history and the evolution of the concept of explainability and its relationships with the legal context in Europe (page 9).

Amelio et al. discuss approaches to interpret and compress convolutional neural networks (CNNs), enhancing their interpretability and efficiency (page 10). Zalewska et al. introduce the BrightBox technology, which provides a surrogate model for interpreting the decisions of black-box classification or regression algorithms (page 12). Spinnato et al. present the LASTS framework, which aims to provide interpretability in black- box time series classifiers (page 14).

Explainable AI in Health Care

The healthcare industry has witnessed the integration of AI systems for various purposes, such as medical imaging analysis, disease diagnosis and personalised treatment. However, the lack of interpretability in AI-based decision-making raises concerns regarding trust and accountability. The authors Bruno et al. highlight the need for addressing the black-box problem by developing an ad-hoc built classifier for lung ultrasound images (page 16). The importance of governing and assessing ethical AI systems in healthcare is emphasised by Briguglio et al. (page 18). Rodis et al. introduce the concept of multimodal explainable AI (MXAI) and its relevance to complex medical applications (page 20). Lädermann et al. discuss the use of machine learning methods in detecting surgical outcomes, aiming to improve patient selection for surgery (page 22). Zervou et al. focus on the application of AI and generative models in precision medicine to streamline the drug discovery process (page 23). These approaches aim to enhance trust, improve patient safety, and provide actionable insights to healthcare professionals.

Explainable AI in Industry

AI plays a crucial role in enhancing productivity and efficiency in industrial applications. However, the lack of explainability in AI models hampers their adoption in critical industrial use cases. Brajovic and Huber focus on integrating AI-specific safety aspects into the automotive development process, particularly addressing the challenges associated with AI application in standard software development (page 24). Folino et al. introduce a ticket-classification framework that integrates deep ensemble methods and AI-based interpretation techniques to support customer support activities (page 27). Jalali et al. propose a counterfactual explanation approach for time-series predictions in industrial use cases, enabling interpretable insights into AI models' decisions (page 28).

Explanations for Chatbots

Generative language models are attracting a lot of attention even in non-technical populations. In many cases, the generated text may not return a faithful representation of truth. Thus, the necessity emerges to provide additional evidence of the elements that are included in the text. Mountantonakis and Tzitzikas present GPT•LODS, a prototype that validates ChatGPT responses using resource description frameworks (RDF) knowledge graphs (page 29). Prasatzakis et al. propose an easy-to-understand and flexible chatbot architecture based on the "event calculus" for high-level reasoning (page 31).

Societal Challenges

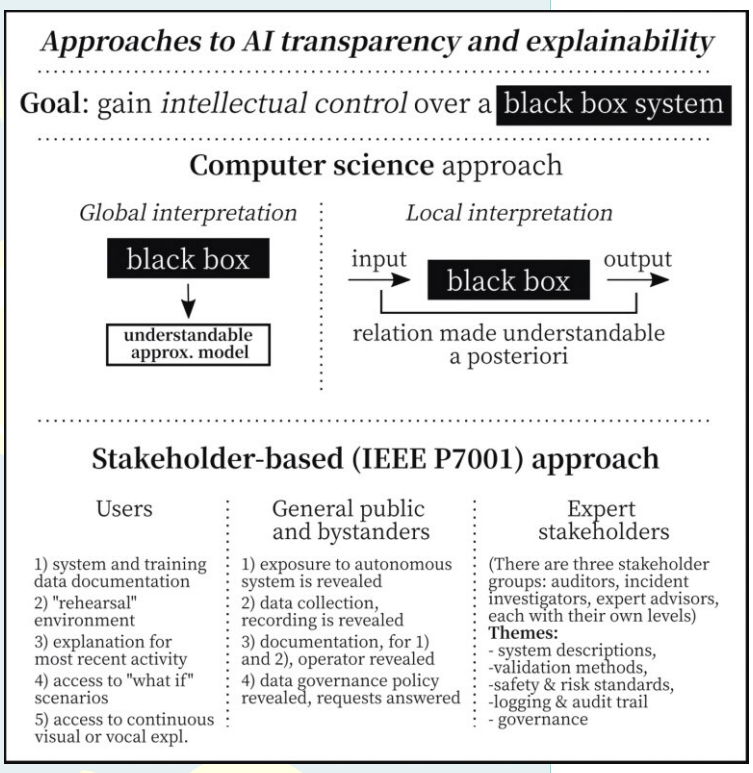
This special issue covers a range of cross-domain applications of XAI that have an impact on several societal challenges, like forest preservation, quality assessment of information and astronomy object detection. It involves developing AI models and systems that can provide transparent and interpretable explanations for their decision-making processes. Jalali and Schindler propose the integration of long short-term memories (LSTMs) with example-based explanations to enhance interpretability in tree-growth models. The aim is to identify critical features impacting outcomes, engage domain experts, address privacy protection, and select appropriate reference models to support informed decision-making in forestry and climate change mitigation (page 33). Ceolin and Qi discuss the design of AI pipelines for automated information quality assessment, which are fully transparent and customisable by end-users. By leveraging reasoning, natural language processing (NLP), and crowdsourcing components, these pipelines enhance transparency, mitigate biases, and aid in the fight against disinformation. Jaziri and Parisot apply XAI techniques to ensure the reliability and absence of bias in deep sky objects classification models used in astronomy.

In conclusion, this special issue showcases several explanation methods and the diverse applications of explainable AI (XAI) across various fields, including healthcare, industry, ethics, climate change, and generative language models. The projects showcased in this issue highlight the importance of transparency and interpretability in complex machine learning models, providing insights into decision-making processes and

empowering stakeholders to understand and trust AI systems. The advancements in XAI contribute to improved diagnostic accuracy, enhanced customer support experiences, ethical AI governance, theoretical developments in model compression and surrogate modelling, interpretability in tree-growth models, integration of AI-specific safety aspects, and combating disinformation. The papers not only provide valuable insights into XAI but also promote further research on XAI, fostering innovation and advancements in understanding AI's internal mechanisms and its impact on various industries.

Please contact:
Manjunatha Veerappa
Fraunhofer Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Germany
manjunatha.veerappa@iosb.fraunhofer.de

Salvo Rinzivillo
CNR-ISTI, Italy
rinzivillo@isti.cnr.it



Explainable AI: A Brief History of the Concept

by Mihály Héder (SZTAKI)

Understandability of computers has been a research topic from the very early days, but more systematically from the 1980s, when human-computer interaction started to take shape. In their book published in 1986, Winograd and Flores [1] extensively dealt with the issues of explanations and transparency. They set out to replace vague terms like "user-friendly", "easy-to-learn" and "self-explaining" with scientifically grounded design principles. They did this by relying on phenomenology and, especially, cognitive science. Their key message was that a system needs to reflect how the user's mental representation of the domain of use is structured. From our current vantage point, almost four decades later, we can see that this was the user-facing variation of a similar idea, but for developers – object-oriented programming, a method on the rise at the time.

The 1980s precedes the now widespread success of machine learning at creating artificial intelligence (AI). In the days of "good old-fashioned" AI, with fewer tools and fewer computational resources, success was built on data structures and logic. These constraints resulted in systems that the creators and adaptors could keep under their intellectual oversight, or at least they knew it was possible to look under the hood and see exactly what was going on.

With machine learning, the designed structures and curated rulesets were replaced by machine-generated models. But, due to the nature of computers, every detail and bit of these models can still be examined easily. This posed a challenge from the terminological point of view: why would we call something a black box (a term Rosenblatt used in the context of artificial neural networks already in 1957, but for a single neuron) if every detail can be readily known? While the word "complexity" is sometimes used – quite confusingly due to its many adjacent meanings – it is more accurate to talk about the lack of understandability or not having adequate explanations about curious behaviour. Understanding is an epistemic value to be achieved by a human investigating a system; therefore, the term "epistemic opacity" [2] was introduced. The opposite of this is then (epistemic) transparency, a feature of a system that affords human understanding and intellectual oversight.

Machine learning, especially deep learning, does not produce models and systems built on these models with this feature, therefore, they create epistemic deficit. Yet, they are here to stay because of their performance. They need to be made transparent, then.

There are two main strategies to interpret, that is, understand these models: first, the entire model may be interpreted, in which case the resulting explanation is called "global" – continuing the tradition of poor choice of terminology in AI (alternatives could have been "comprehensive", "broad", etc.).

This can be achieved by a surrogate system, which helps by faithfully representing the original model while allowing for simplification, and uses elements that humans easily understand. If such a surrogate is successfully made, the entire model is made transparent. Moreover, we can predict its behaviour to imagined inputs before it happens, providing us with intellectual control. Other global explanations visualise the model or map out concepts used by a model. We can only speculate regarding the etymology, but most probably, this approach is called "global" because it is the apparent linguistic opposite of "local". This word takes us to the second interpretation strategy, local interpretation. The usage of "local" is much better justified by the concept of local fidelity – it means that an explanation is made for one particular output of a system, but in a way that it may be used for similar inputs, where similarity is measured as the distance in a mathematical space. Therefore, we are talking here about true spatial locality.

This epistemic approach to transparency is inevitably relative to the knowledge of the particular persons trying to achieve intellectual oversight. This fact is best engaged by the IEEE P7001 standard draft [3], which is expected to become a harmonized EU standard as a part of the EU AI Act; legislation that makes transparency (and therefore explainability) central.

This approved draft uses a stakeholder-based approach and divides humans into "users", the "general public" or "bystanders" (non-users who may still be affected), and "experts". The last group is further divided into certification agencies and auditors, incident investigators and expert advisers in litigation. This draft is very helpful, as it clarifies that the mathematical method under the XAI umbrella term is for the experts, while user transparency may be created by layperson explanations, like for clustering the term "other users who listened to this also liked the following". Agencies are catered for yet another, more administrative modus of transparency, tuned for accountability.

As transparency is widely believed to be essential to build trust, the methods to achieve it are here to stay, and therefore explainable AI has a long future.

References:

[1] T. Winograd, F. Flores and F. F.Flores, Understanding Computers and Cognition: A New Foundation for Design, Intellect Books, 1986.
[2] M. Héder, "The epistemic opacity of autonomous systems and the ethical consequences," AI & Society, 1–9, 2020.
[3] A. F. T. Winfield, et al., "IEEE P7001: A proposed standard on transparency," Frontiers in Robotics and AI, vol. 8, 665729, 2021.

Please contact:
Mihály Héder, SZTAKI, Hungary
mihaly.heder@sztaki.hu

A Multilayer Network-Based Approach for Interpreting and Compressing Convolutional Neural Networks

by Alessia Amelio, Gianluca Bonifazi, Domenico Ursino and Luca Virgili (Polytechnic University of Marche)

We propose an approach to map a convolutional neural network (CNN) into a multilayer network. It allows the interpretability of the internal structure of deep learning architectures. Then, we use this representation to compress the CNN.

Researchers have recently become more aware of the necessity to scale back the size and complexity of deep neural networks. As a result, a number of techniques are being suggested to shrink the size of current networks without significantly impacting their performance. Exploring the many layers and components of a deep learning model is crucial in order to achieve this goal. In fact, one could pinpoint the most important components, the most relevant patterns and features, the information flow and so on. We want to make a contribution in this setting by proposing a new way of interpreting and exploring a CNN through a multilayer network representation of it, which is then used for compressing it [1].

We operate under the assumption that deep learning networks may be represented, analysed, explored and otherwise greatly supported by complex networks, particularly multilayer ones. Accordingly, we first introduce a method to transform deep learning networks into multilayer ones and then exploit the latter to explore and manipulate the former. Our study focuses on the CNN, which is a specific type of deep learning network widely adopted in different fields, especially computer vision; however, it can easily be extended to other kinds of deep learning networks. The multilayer network is a particular graph-based data structure composed of different layers. Each layer represents a graph with a specific type of connection among the nodes. Multilayer networks are a type of complex networks sophisticated enough to represent all the main components of a CNN. In fact, all the typical elements of a CNN (i.e. nodes, connections, filters, weights, etc.) can be represented through the basic components of a multilayer network (i.e. nodes, arcs, weights and layers). Once the representation of the CNN by the multilayer network has been obtained, the latter is adopted to explore and manipulate the former. To prove its potential, we use this representation to provide a method for removing unnecessary convolutional layers from a CNN. This method looks for layers in the CNN that can be pruned without significantly affecting the CNN performance and, if it finds any, it goes ahead and removes those layers, returning a new CNN [1].

More specifically, mapping the CNN into a multilayer network is performed in different steps (see Figure 1). In the first step, the CNN is trained from a database of images labelled with

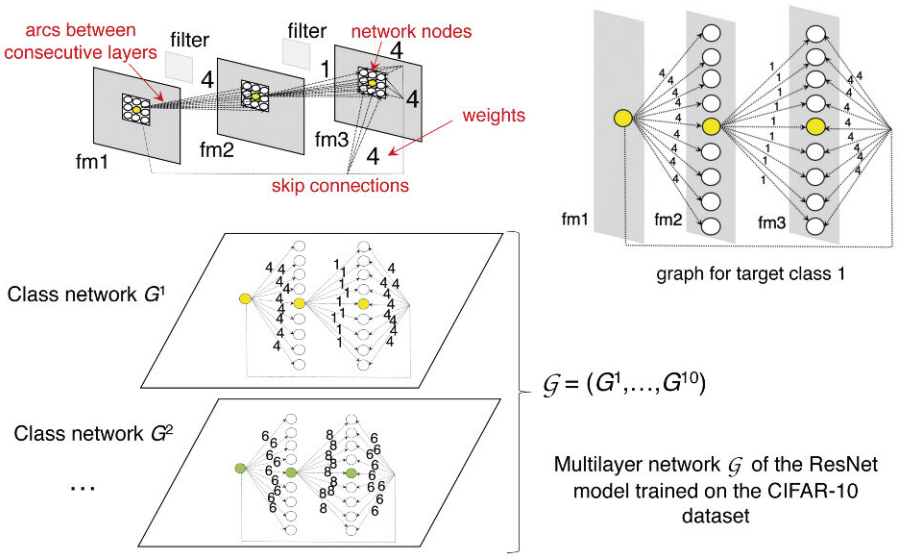


Figure 1: Mapping a Residual Neural Network (ResNet) into a multilayer network for the CIFAR-10 dataset [3]. Three convolutional layers of the ResNet produce three feature maps (fm1, fm2, fm3). Nodes of the multilayer network correspond to elements of the three feature maps. Arcs are created between adjacent nodes of subsequent feature maps. The weight of an arc between two nodes corresponds to the activation value of the first node in its feature map. There is one layer for each target class of the dataset, resulting in a multilayer network with ten layers (class networks).

different target classes. Each image of the training set is forwarded through the network in order to predict its class. Then, for that image, the classification error between the predicted and target class is computed and back-propagated through the network for tuning its parameters. In the second step, each set of images of the dataset belonging to a target class is provided as input to the CNN in order to create a layer for the multilayer network. For each feature map of the CNN, each element becomes a node of the layer, and is linked to other nodes derived from the next feature maps according to their spatial adjacency. The weight of the arc between two nodes corresponds to the activation value associated with the first node in its feature map [2,3].

We start from the assumption that nodes of the multilayer network with higher degree (defined for a node as the number of arcs crossing it) correspond to more informative areas of the feature maps of the CNN. Accordingly, the CNN is compressed through the following steps (see Figure 2). First, the degree of each node in each layer of the multilayer network is computed. Then, the total degree of each node over the different layers is calculated. Afterward, nodes with a total degree higher than a threshold are detected. In particular, the latter is computed as the mean degree of all nodes multiplied by a scaling factor. Finally, the only feature maps containing selected nodes are retained in the CNN, while the other ones are removed [2,3].

We adopted our approach for compressing two well-known CNNs, i.e. VGG [2] and ResNet [3], on different benchmark datasets in computer vision. The adopted datasets are: (i) MNIST, for the identification of handwritten digits from 0 to 9; (ii) CALTECH-101, for the recognition of objects belonging to 101 distinct classes; (iii) CIFAR-10; and (iv) CIFAR-100, for the recognition of objects belonging to 10 and 100 distinct classes, respectively. The obtained results show that

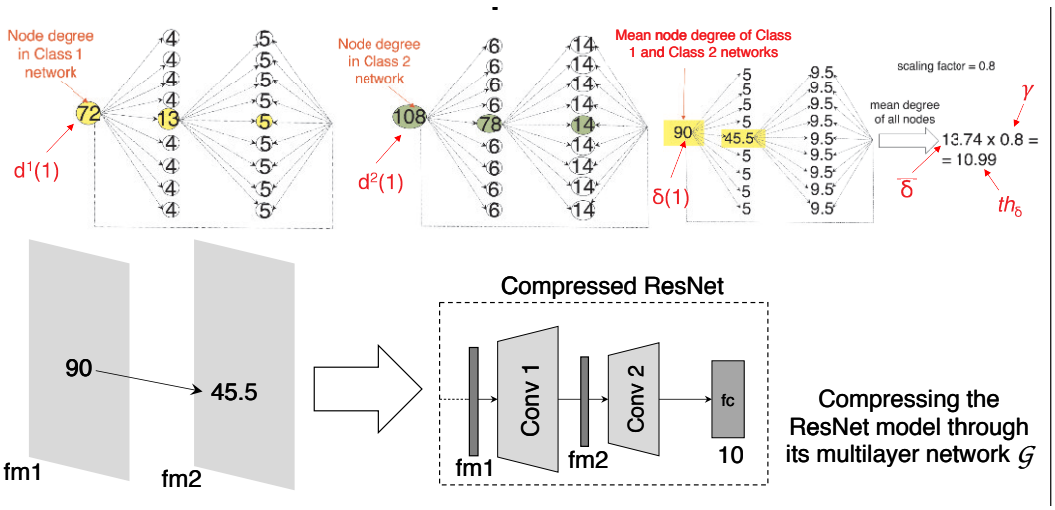


Figure 2: Compressing the ResNet of Figure 1 when two target classes are considered [3]. The two layers of the multilayer network corresponding to the target classes are reported on the top left and top mid of the figure (yellow and green coloured). The third graph on the top right contains the total degree of each node over the two layers. The value 10.99 of the threshold is obtained as the mean degree of all nodes, which is 13.74, multiplied by the scaling factor, which is 0.8. The only two nodes exceeding the threshold are those with mean degrees of 90 and 45.5. Since they are located on the first and second feature maps, fm1 and fm2, only the first and second convolutional layers will be retained in the CNN.

our approach overcomes, in terms of different performance measures, another similar approach that uses a single-layer network for representing the CNN, as well as other approaches for compressing CNNs [2,3] proposed in the past literature.

This paper should not be considered as an endpoint but rather as a starting point for further research. A possible advancement in this direction is the development of a mechanism that allows the visualisation of the network to facilitate the possible interpretation of the compression result.

The approach described in this paper is the result of a collaboration between the Department of Engineering and Geology, University "G. d'Annunzio" Chieti-Pescara, Italy, and the Department of Information Engineering, Polytechnic University of Marche, Italy. A GitHub repository with the source code of the proposed approach is available at [L1].

Link:
[L1] https://github.com/lucav48/cnn2multilayer

References:
[1] A. Amelio, G. Bonifazi, E. Corradini, et al., "Mapping and compressing a convolutional neural network through a multilayer network," presented at the 30th Symposium on Advanced Database System, Tirrenia (Pisa), Italy, June 19–22, 2022.
[2] A. Amelio, G. Bonifazi, E. Corradini, et al., "A multilayer network-based approach to represent, explore and handle convolutional neural networks," Cognitive Computation, vol. 15, pp. 61–89, 2023.
[3] A. Amelio, G. Bonifazi, F. Cauteruccio, et al., "Representation and compression of Residual Neural Networks through a multilayer network based approach," Expert Systems with Applications, vol. 215, 119391, 2023.

Please contact:
Luca Virgili, DII, Polytechnic University of Marche, Italy
luca.virgili@staff.univpm.it

Alessia Amelio, InGeo, University "G. d'Annunzio" Chieti-Pescara, Italy
alessia.amelio@unich.it

Diagnostic Explainability by BrightBox

by Andżelika Zalewska-Küpçü (QED Software), Andrzej Janusz (University of Warsaw & QED Software) and Dominik Ślęzak (University of Warsaw & QED Software)

BrightBox technology presents a novel approach to investigating mistakes in machine learning model operations.

The main feature offered by BrightBox technology [1] is the capability to create a surrogate model that can closely imitate any black-box classification or regression algorithm. The approximations of the diagnosed model's predictions are computed using an ensemble of approximate reducts, which are irreducible subsets of attributes preserving almost the same information about the decisions as the whole set of attributes. The prediction is made separately for each reduct, and the results are then averaged. It is worth noting that the decision values for the surrogate are the predictions made by the diagnosed model, not the true target values. An ensemble of approximate reducts can then be used as a classification model.

BrightBox technology is based on neighbourhoods. The surrogate models are used to identify neighbourhoods of instances that have been processed by a machine learning model. The neighbourhood for a diagnosed instance relative to a single reduct is a subset of instances from the reference dataset which belong to the same indiscernibility class. The final neighbourhood is the sum of neighbourhoods computed for all reducts in the ensemble. The neighbourhoods consist of historical instances that were processed in a similar way by rough set-based models.

The neighbourhood is analysed for consistency in labels (ground-truth labels, original model predictions, and their approximations), its size, and the uncertainty of predictions. Such features are called diagnostic attributes. By analysing the mistakes made in these neighbourhoods, we can gain valuable insights into the reasons behind the poor performance of machine learning models.

BrightBox is diagnosing black-box models without requiring any knowledge about the model or direct access to it. The diagnosis is performed on the diagnosed dataset and the model's outputs, like predictions or classifications, without using predictions on the training set. This approach is particularly useful in industrial environments where machine learning models are deployed.

To construct a surrogate model in the current version of BrightBox, it is necessary that all attributes have discrete values. Therefore, the first step of the approximation procedure involves discretising all numeric attributes using the quantile method.

To construct the approximator, values for two hyper-parameters need to be selected: epsilon, which represents the threshold for reducts approximation, and the number of reducts in the ensemble. Since the goal is to find the best possible

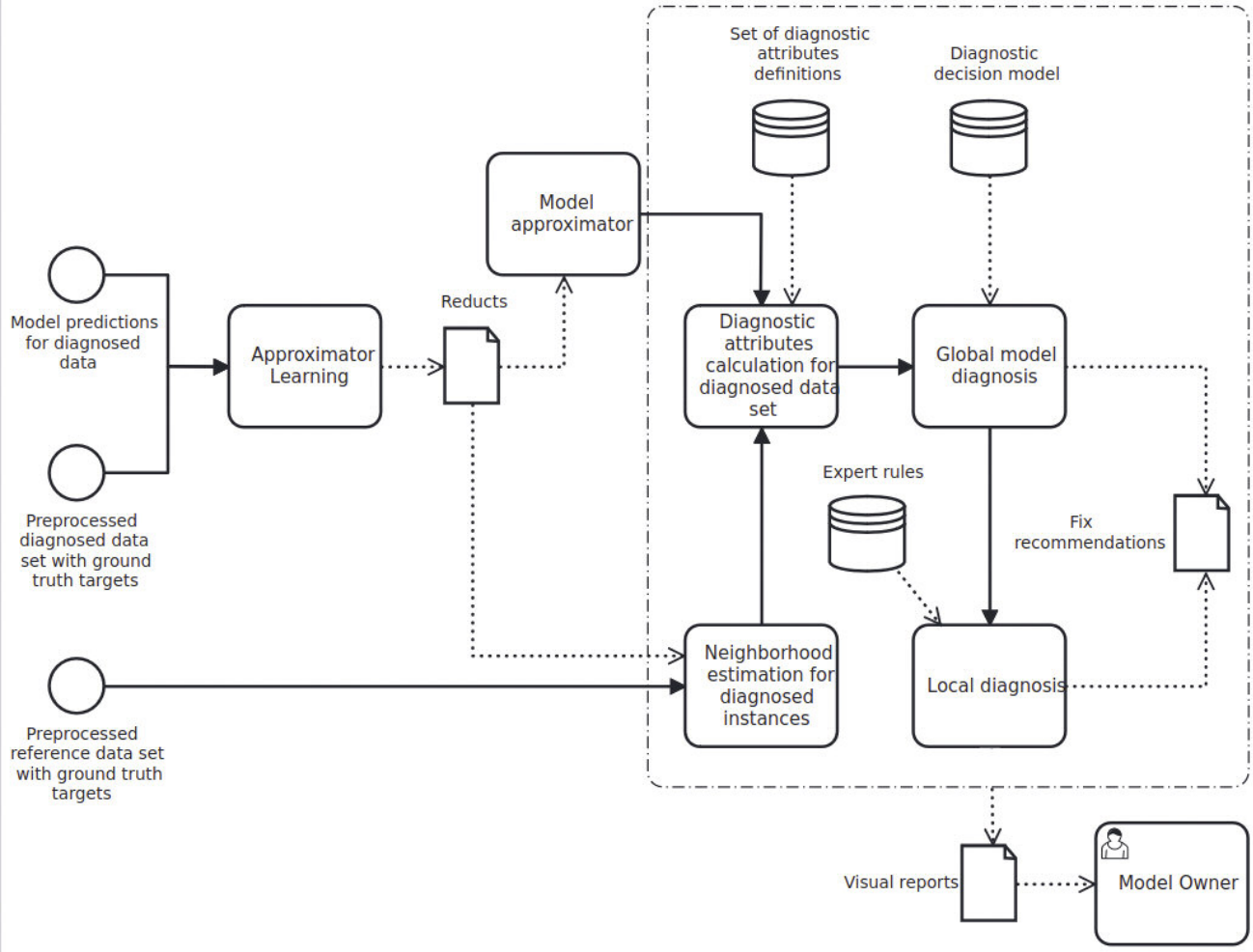


Figure 1: Diagram of diagnostic method workflow.

approximation of the model's predictions, a grid search is performed to tune these hyper-parameters. The final selection of the surrogate model is based on achieving a minimum Cohen's kappa value of 0.9 to ensure high-quality approximation. If this quality cannot be attained, the ensemble of reducts is trained with the settings that provide the best-possible approximation quality. The resulting approximator is then utilised to determine neighbourhoods of instances from the diagnosed data and to compute values of diagnostic attributes.

Next, in the BrightBox diagnostics process, a global diagnosis of the model is conducted. This involves computing a summary of diagnostic attribute values obtained for the diagnosed dataset, followed by the use of a pretrained classifier to categorise the investigated model into one of three categories: Near-optimal fit, Under-fitted model, or Over-fitted model. The classifier is pretrained on a manually-labeled dataset consisting of diagnostic attribute summaries computed for numerous datasets and commonly used prediction models.

The next step involves examining individual data instances and predictions made by the diagnosed model. This analysis utilises the diagnostic attribute values computed earlier, along with the output of the global diagnostic model, to provide a comprehensive assessment of model predictions and related issues. Moreover, a set of local diagnostic rules, designed by experts, is applied to provide users with accessible insights. An example of such a rule and its corresponding fix recommendation is: "If the model is diagnosed as over-fitted and the incorrectly classified instance has low model uncertainty, and its neighbourhood is not small, then the error is likely due to overfitting. Improve the model fitting procedure."

After completing the diagnosis process, BrightBox generates visual reports emphasising the most important findings. They include relevant statistics related to the diagnosed model, the quality of its approximator, and the distributions of diagnostic attributes. Interactive plots help to explore diagnoses for individual instances and analyse their statistics for specific groups. Reports also offer valuable insights on the importance of original attributes, approximated by the significance of attributes in the surrogate model. This diagnostic process is both efficient and effective in providing us with a deeper understanding of model operations. Figure 1 shows the entire workflow of BrightBox.

In conclusion, BrightBox employs ensembles of rough set-based reducts to approximate black-box machine learning models. This approach is at the core of XAI, which aims to make black-box models more interpretable. However, in our case, these ensembles also enable us to identify collections of historical instances that are processed in a similar way to each new instance. These collections, known as neighbourhoods, help us categorise errors made by diagnosed models.

Actually, analysing neighbourhoods – rather than just attribute values – is the key advantage of BrightBox. This means that diagnostic attributes can reveal valuable characteristics of specific data instances, providing domain experts and data scientists with the needed information to create better prediction models. Focusing on neighbourhoods, the technology offers deeper insights and more meaningful guidance for model improvement.

Our goal is to further develop BrightBox in practical applications (see e.g. [2]). This approach can be highly beneficial in, for example, detecting errors in models submitted for online data science competitions, allowing their organisers and sponsors to gain valuable insights into the performance of each solution. Such insights can then suggest improvements to the models and facilitate their deployment in production-ready environments. In addition, it is possible to construct improved solutions from the existing ones. For instance, by using information about different degrees of risk (uncertainty) of decision-making by different classifiers of an ensemble, it is possible to modify the procedure for resolving conflicts in voting.

References:

[1] A. Janusz, et al., "BrightBox – a rough set based technology for diagnosing mistakes of machine learning models," Applied Soft Computing, vol. 141, pp. 110285, 2023. https://doi.org/10.1016/j.asoc.2023.110285

[2] A. Janusz and D. Ślęzak, "KnowledgePit meets BrightBox: a step toward insightful investigation of the results of data science competitions," in Proc. FedCSIS 2023 in ACSIS vol. 30, pp. 393-398. https://doi.org/10.15439/2022F309

Please contact:

Andżelika Zalewska-Küpçü,
QED Software, Poland
andzelika.zalewska-kupcu@qed.pl

Dominik Ślęzak,
University of Warsaw & QED Software, Poland
slezak@mimuw.edu.pl

An Explanation that LASTS: Understanding Any Time Series Classifier

by Francesco Spinnato (Scuola Normale Superiore and CNR-ISTI), Riccardo Guidotti (University of Pisa) and Anna Monreale (University of Pisa)

We present LASTS, an XAI framework that addresses the lack of explainability in black-box time series classifiers. LASTS utilises saliency maps, instance-based explanations and rule-based explanations to provide interpretable insights into the predictions made by these classifiers. LASTS aims to bridge the gap between accuracy and explainability, specifically in critical domains.

In recent years, the availability of high-dimensional time series data has led to the widespread usage of time series classifiers in various domains, including health care and finance. These classifiers play a crucial role in applications such as anomaly detection in stock markets and the automated diagnosis of heart diseases.

The existing landscape of time series classifiers encompasses a range of approaches. However, despite their effectiveness in achieving high classification accuracy, most of these classifiers suffer from a critical limitation: they are black-box models, offering little insight into their decision-making process [1].

The lack of explainability in black-box time series classifiers poses challenges, particularly in critical domains where human experts must understand the reasons behind the model's predictions. In applications such as clinical diagnosis, the interpretability of the models used by artificial intelligence (AI) systems becomes essential for building trust and facilitating reliable interaction between machines and human experts. Meaningful explanations can enhance the cognitive ability of domain experts, allowing them to make informed decisions and supporting AI accountability and responsibility in the decision-making process [L1].

We propose the LASTS (Local Agnostic Subsequence-based Time Series Explainer) framework to address the need for explainability in time series classification, providing interpretable explanations for any black-box predictor. By unveiling the logic behind the decisions made by these classifiers, LASTS enhances transparency and facilitates a deeper understanding of the classification process. The first version of LASTS was published in [2]. Since then, we have made significant advancements, introducing heterogeneous explanations and a novel salience map extraction to explain both univariate and multivariate time series. Compared to the previous version, these enhancements provide a more comprehensive, interpretable, and versatile approach to explaining black-box time series classifiers.

The input to the LASTS framework is a time series, X, and a black-box classifier, while the output is an explanation for the black-box's decision. The explanation comprises three parts: a

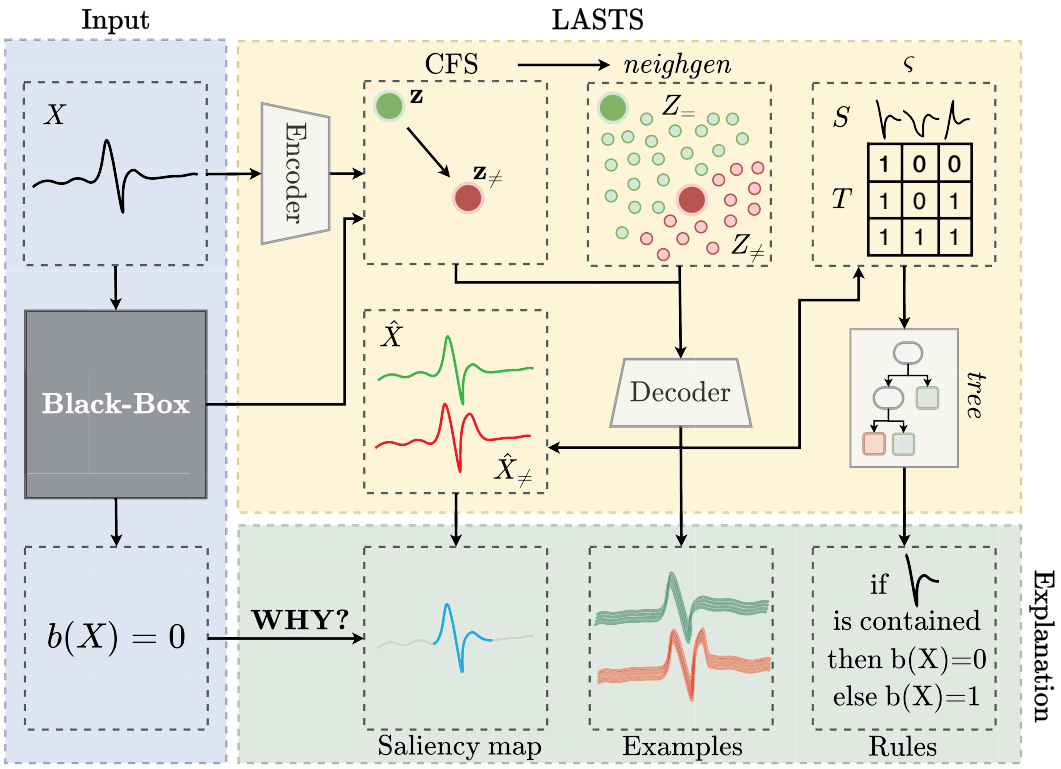


Figure 1: A comprehensive schema of LASTS.

saliency map, an instance-based explanation, and a rule-based explanation (a schema is presented in Figure 1). The saliency map highlights the most influential parts of the time series that contribute to the classifier's decision. This visualisation provides an immediate assessment of the critical timesteps, enabling users to better understand the driving factors behind the classification outcome. The instance-based explanation employs a set of exemplar and counterexemplar time series, offering concrete examples that align with or diverge from the black-box classifier's decision. These instances help identify common patterns and highlight the necessary modifications for obtaining different classification outcomes. Finally, the rule-based explanation utilises logical conditions based on interpretable time series subsequences, providing factual and counterfactual rules that reveal the reasons for the classification. The novel component of LASTS is its integration of multiple explanation components, which provide a comprehensive and interpretable set of explanations, that can be useful for different kinds of users.

From a technical standpoint, the framework leverages a trained variational autoencoder to encode and decode time series into a latent space. Once the input time series is encoded into its latent representation, LASTS finds the closest counterexemplar using a novel search algorithm, generating a synthetic neighbourhood around the black-box's decision boundary. The neighbourhood is then decoded, obtaining the black-box predictions for these synthetic instances. The saliency map is extracted based on the distance between X and the closest decoded counterexemplar, while other exemplar and counterexemplar instances are derived from the neighbourhood. Lastly, the neighbourhood is transformed into a set of subsequences, and a decision-tree surrogate is trained on the transformed data to extract the factual and counterfactual rules.

An example explanation for a real electrocardiogram from the ECG5000 dataset [L2] is presented in Figure 2. The black box classifies the instance being explained as a "normal" heartbeat. The explanation for this prediction highlights the main differ-

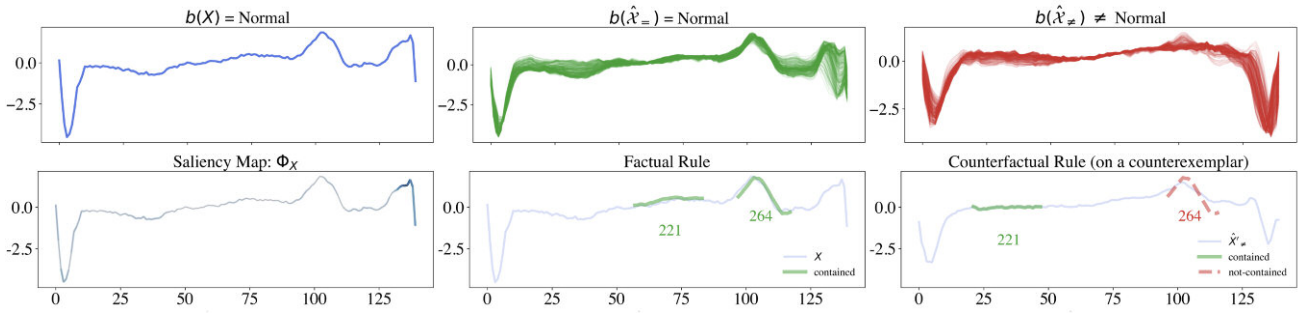


Figure 2: Explanation of a black-box prediction for a heartbeat of the ECG5000 dataset. From left to right: (top) instance to explain, exemplars, counterexemplars, (bottom) saliency map, factual rule and counterfactual rule (shown over a counterexemplar).

ence between normal and abnormal heartbeats: the lower and evident V-shape in the rightmost part of the time series. The saliency map confirms this observation by highlighting the last observations of the time series. The rules further illustrate the differences between classes, with the factual rule indicating the presence of a specific subsequence in normal instances and the counterfactual rule indicating its absence in abnormal instances. While the saliency map and rules may not cover the exact same areas, they provide complementary insights into the discriminative features of the time series.

Overall, LASTS represents a significant advancement in the field of time series explanation, with promising potential for future research and practical applications. In our future research, we plan to explore several directions to enhance LASTS. Firstly, we intend to evaluate the framework on longer and more complex real-world time series datasets, aiming to validate its performance in challenging scenarios. Additionally, we aim to extend LASTS to other types of sequential data, such as trajectories, text, and shopping transactions, in order to broaden its scope of applicability. Secondly, we will delve deeper into the relationship between the latent and subsequence spaces, conducting further investigations to gain a comprehensive understanding of their interactions. Finally, we intend to conduct human decision-making tasks guided by LASTS explanations, offering practical evaluation and valuable insights into the effectiveness of the explanations in real-world decision scenarios.

This article is coauthored with Mirco Nanni, Fosca Giannotti, and Dino Pedreschi (CNR-ISTI, Scuola Normale Superiore, Università di Pisa).

Links:
[L1] https://artificialintelligenceact.eu/
[L2] https://kwz.me/hxK

References:
[1] A. Theissler, et al., "Explainable AI for time series classification: a review, taxonomy and research directions," IEEE Access, 2022.
[2] R. Guidotti, et al., "Explaining any time series classifier," 2020 IEEE 2nd Int. Conf. on Cognitive Machine Intelligence (CogMI), IEEE, 2020.

Please contact:
Francesco Spinnato
Scuola Normale Superiore and CNR-ISTI, Pisa, Italy
francesco.spinnato@sns.it
Riccardo Guidotti
University of Pisa, Pisa, Italy
riccardo.guidotti@unipi.it

Explaining Ensemble Models for Lung Ultrasound Classification

by Antonio Bruno, Giacomo Ignesti and Massimo Martinelli (CNR-ISTI)

Correct classification is the main aspect in evaluating the quality of an artificial intelligence system, but what happens when you reach top accuracy and no method explains how it works? In our study, we aim at addressing the black-box problem using an ad-hoc built classifier for lung ultrasound images.

In the last few years, the novelties of artificial intelligence (AI) and computer vision (CV) significantly increased, allowing new algorithms to obtain meaningful information from digital images. Medicine is a field in which the use of this technology is experiencing fast growth. In 2020, in the USA alone, the production of 600 million medical images was reported, and this number seems to increase steadily. Robust and trustworthy algorithms need to be developed in a multidisciplinary collaboration.

During the SARS-CoV-2 pandemic, a fast and safe response became even more necessary. The use of point-of-care ultrasound (POCUS) to detect SARS-CoV-2 (viral) pneumonia and the bacterial infection emerged as one of the most peculiar emerging case studies, involving the use of on-site ultrasound examinations rather than a dedicated facility. As well as being faster, safer and less expensive, lung ultrasound (LUS) also appears to detect signs of lung diseases as well as or even better than other methods, such as X-ray and computed tomography (CT).

The employment of lung POCUS seemed an optimal solution for both quarantined and hospitalised subjects. CT and magnetic resonance imaging (MRI) are far more precise and reliable examinations, but both have downsides over mass screening. In our study, an efficient adaptive minimal ensembling model was developed to classify LUS using the largest publicly available dataset, the COVID-19 lung ultrasound dataset [L1], composed of 261 ultrasound videos and images from 216 different patients. The General Data Protection Regulation

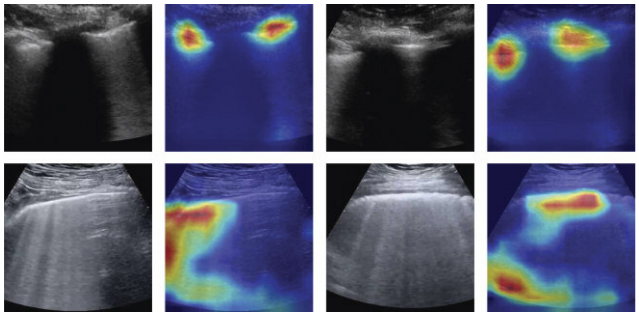


Figure 1: COVID-19 – Original and Grad-CAM-processed samples are shown for subjects with COVID-19; different images within different subjects show similar activation maps.

(GDPR) and the European Committee AI Act focus on the intent that an AI-deployed system should be trustworthy and fully explainable.

Several explainability approaches arose from the scientific community and new ones are under development. Focusing on image interpretation, a debate about which approach should be used is ongoing. As the core model, we selected EfficientNet-b0 [1] because of its good accuracy/complexity trade-off.

The efficiency of this architecture is given by two main factors:

- the reduced number of parameters given by compound scaling, by which input scaling (i.e. input size), width scaling (i.e. convolutional kernel size) and depth scaling (i.e. number of layers) are performed in conjunction because they are dependent
- the low number of FLOPs (floating point operations) of the inverted bottleneck MBConv (first introduced in MobileNetV2, an efficient model designed to run on smartphones) as a main constituent block.

	Class	Recall	Precision	F1-Score
InceptionV3				
Acc.: 89.1% (89.3%)	COVID-19	0.864±0.036	0.901±0.031	0.880±0.030
#Param.: 23.9M	Pneumonia	0.908±0.025	0.842±0.037	0.871±0.025
FLOPs: 6G	Healthy	0.907±0.026	0.918±0.021	0.911±0.021
DenseNet-201				
Acc.: 90.4%	COVID-19	0.892	0.918	0.905
#Param.: 20M	Pneumonia	0.903	0.610	0.728
FLOPs: 4.29G	Healthy	0.850	0.842	0.846
Light Transformer				
Acc.: 93.4%	COVID-19	0.958±0.025	0.958±0.012	0.951±0.017
#Param.: 0.3M	Pneumonia	0.948±0.013	0.951±0.038	0.949±0.020
FLOPs: 1G	Healthy	0.877±0.034	0.912±0.037	0.894±0.036
Weak model (our)				
Acc.: 98.7% (98.3%)	COVID-19	0.984±0.004	0.993±0.004	0.990±0.004
#Param.: 5M	Pneumonia	0.997±0.005	0.991±0.006	0.991±0.007
FLOPs: 0.39G	Healthy	0.999±0.003	0.993±0.003	0.995±0.004
Ensemble (our)				
Acc.: 100% (100%)	COVID-19	1.000±0.000	1.000±0.000	1.000±0.000

Table 1: Test set comparisons (on 5-fold cross-validation), with metrics for each class, of the proposed model with the SOTA. Our solution outperforms the SOTA on all metrics and has the lowest complexity.

The greatest contribution of our study is given by the introduction of an ensembling strategy. Due to its resource-consuming nature and the exponential growth of model complexity, ensembling is scarcely used in computer vision, but we demonstrate how to perform it in an adaptive and efficient way:

- using only two weak models (minimality, efficiency)
- performing the ensemble using a linear combination layer, trainable by gradient descent (adaptivity)
- performing it using the features instead of the output, excluding redundant operations (efficiency)
- fine-tuning the combination layer only (efficiency).

The linear combination layer used to perform the adaptive ensemble can be described by the following equation:

$$Feat_{comb} = W_{comb} Feat_{concat} + b_{comb} \ ,$$

where:

- $Feat_{concat} = feat_{weak1} \cdot feat_{weak2}$ is the concatenation of the weak features
- $feat_{weak1}$, $feat_{weak2}$ are the features obtained by the two weak models.

Table 1 shows that the ensemble further reduces the variance and improves the generalisation power (i.e. performance on validation and test dataset), outperforming the state-of-the-art (SOTA) with lower complexity (moreover, the complexity of this ensemble strategy is equal to the complexity of a single EfficientNet-b0, since the processing of the weak models is independent and parallelised).

Even if our model can give extremely accurate and fast responses, it is crucial that it also is secure and understandable. To this aim, we applied Grad-CAM, since this method uses gradients with respect to a particular convolutional feature map (in our case the last convolutional layer of the model) to identify the regions on the input that are more discriminative for the classification results (i.e. higher gradient value means higher contribution to the classification).

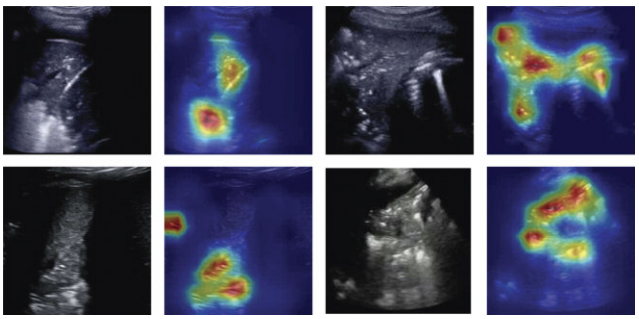


Figure 2: Pneumonia – Original and Grad-CAM-processed samples are shown for subjects with pneumonia; the attention of the classifier is on different regions of the images in contrast to the other two classes.

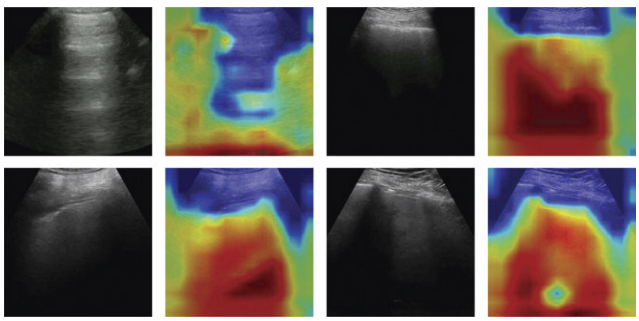


Figure 3: Healthy – Original and Grad-CAM-processed samples are shown for healthy subjects; the classifier focuses its attention all over the image or outside; it seems it does not find relevant information, unlike the cases with a pathology.

The application of this technique to our model gives reasonable results. A first non-trivial result is that for the images classified in the same way, a similar gradient is activated, which in turn originates from similar areas in the images. This result was further investigated by comparing the saliency map of the same correctly classified images between the EfficientNet-b0 and the ensemble models. While both identify similar parts, the ensemble model seems to maintain attention on a more concentrated area.

Figures 1, 2 and 3 show that there are typical signs of evidence for each class:

- COVID-19 (Figure 1) – usually, more concentrated and relatively large, mainly over the pleural line and on the "edges"
- Pneumonia (Figure 2) – evidence mainly below the pleural line, with widespread area having spots
- Healthy (Figure 3) – mainly the very expanded, healthy part of the lung (black).

Even if this study seems to provide robust and interpretable results, it lacks in-depth research on the effective stability of the explanation. To further test our method, in an ongoing telemedicine project [2], in close collaboration with specialist sonographer staff, we are going to use further explainability methods on other important signs that can be present in LUS (e.g. A-lines, B-lines, thickness), in order to improve the correlation with the ground truth.

Link:

[L1] https://github.com/jannisborn/covid19_ultrasound/blob/master/data/README.md

References:

[1] Mingxing Tan and Quoc V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114, 2019. Available at: http://proceedings.mlr.press/v97/tan19a.html

[2] G. Ignesti et al., "An intelligent platform of services based on multimedia understanding and telehealth for supporting the management of SARS-CoV-2 multi-pathological patients," in Proc. 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2022, pp. 553–560. Available at: https://doi.org/10.1109/SITIS57111.2022.00089

Please contact:

Massimo Martinelli, CNR-ISTI, Italy
massimo.martinelli@isti.cnr.it

A Governance and Assessment Model for Ethical Artificial Intelligence in Healthcare

by Luigi Briguglio, Francesca Morpurgo and Carmela Occhipinti (CyberEthics Lab.)

How can clinicians be deemed responsible for basing their decisions on diagnoses generated by artificial intelligence and derived in a way that cannot be fully understood? How can patients rely and accept decisions if they are based on "black boxes" of data and algorithms? In the context of the MES-CoBraD project, CyberEthics Lab. defines a model for governing and assessing "Ethical Artificial Intelligence" (ETHAI).

Innovative technologies and approaches have enabled the evolution of many sectors and the increased well-being of our society, reducing time to produce and make available solutions to the public, and improving quality of life. Healthcare services have been a core part of this process, achieving impressive milestones in less than a hundred years. The COVID-19 pandemic has shown how reliable scientific research can be, producing vaccines in a relatively short time and under difficult working conditions [1].

Successful clinical care requires that all people have access to good-quality health care and that they can comply with recommended treatments. However, the ability to respond to and support the health demands of citizens depends on both the health system of a country and the economic and health structure of society (e.g. an ever-growing rate of ageing population means more attention).

This reflects in the global trend of innovating health care by introducing digitised and distributed "neighbourhood healthcare centres" capable of offering wider access to care for citizens. Delivering better health care means a more efficient and effective intervention in diagnosis, and thus prevention and early detection of non-communicable diseases can be better tracked. Ensuring health and promoting well-being for all and for all ages is set out as the Goal 3 of the 2030 Agenda for Sustainable Development of the United Nations [2].

In the digital era, healthcare (eHealth) and prevention actions will be characterised by: (i) preventive health care based on diffused remote monitoring through connected devices, that is, the Internet-of-Things (IoT) and predictive disease detection through computing capabilities, that is, Artificial Intelligence (AI); (ii) prevention actions, continuous contact with healthcare providers and smarter medication; (iii) remote and in-hospital assistance and surgery with dedicated robotics; (iv) asset- and intervention-management in hospitals and healthcare centres [3].

In this context, researchers of the Multidisciplinary Expert System for the Assessment & Management of Complex Brain Disorders (MES-CoBraD) project [L1] are working together to exploit the potential of data and AI to develop a common innovative protocol for the accurate diagnosis and personalised

care of complex brain diseases, with a primary focus on improving the quality of life of patients, their caregivers, and the society at large.

To this aim, MES-CoBraD is developing an eHealth expert system that can be used during the diagnostic process to support clinicians' decision making, analyse data and give access to a huge knowledge base through a secure data lake, where all data gathered in the hospitals participating in the research are placed.

Thanks to the fact that multidisciplinary teams are involved, MES-CoBraD is not only addressing the technological aspects that regard the definition of algorithms and data, but is also considering how to embed ethics, legal and social dimensions into this Multidisciplinary Expert System (MES).

Moving from the definition of a common framework of ethics as applied to AI and to the treatment through it of complex brain diseases, the concern here is twofold:

1. Designing and developing a system that will behave "ethically" and that will not deliver biased results at scale or behave as a black box. This implies the necessity to understand and elicit requirements to be followed by the developers of the system, as well as to specify and implement the system in compliance with ethics principles (e.g. transparency and explainability)
2. Helping the clinician, user of this system, to reach a decision that – based on the outcomes of the system – is ethically sound. In this context, the expert system may influence the prognosis of the clinician, impact the relationship between the clinician and patient, and impact the accountability and civil liability of the clinician. In this case, an ethical decision model has to be developed, discussed with all the involved stakeholders, and adopted.

A set of ethical principles lay the foundation of requirements and the ethical decision model, and therefore the decision made by users of the expert system. The usual principles of bioethics (i.e. beneficence, non-maleficence, justice and autonomy) are the most widely used when dealing with any such models. However, these leave out a really important aspect of the latest approaches to health care, that, incorporating some inputs and hints coming from the feminist theoretical reflection, underline the importance – especially in the context of health systems – of the principle of care, adopting a more "humanistic" stance.

For this reason, in the Ethical Artificial Intelligence (ETHAI) model that MES-CoBraD researchers are developing, care is one of the most relevant principles. It lies at the heart of the system together with the other four principles. Bringing medicine back to its foundations, even when technologically enhanced in such a strong way, requires the respect and care of anything that is human (see Figure 1).

Beyond these respectful principles, it is also important to consider the lawful basis of regulations and standards representing the governance layer for this disruptive technology. Indeed, following the "EU Ethics Guidelines for Trustworthy AI" (2018), the European Commission unveiled a proposal for a new Artificial Intelligence Act (AI Act) in April 2021. On May 11 2023, the European Parliament adopted a draft negotiating mandate. This draft of the AI Act includes obligations for

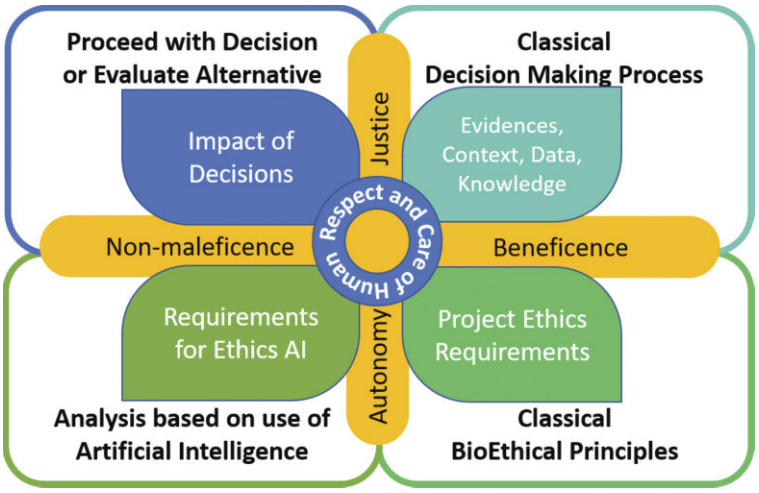


Figure 1: Graphical representation of the ETHAI model.

providers of AI foundation models who would have to guarantee robust protection of fundamental rights, safety and rule of law. Therefore, ethics and legal assessment frameworks, including methodologies and tools, will lay the foundation for any future AI development in the next years, after the entry in force of the AI Act, in order to assess and mitigate risks and comply with design, information and environmental requirements. Ethics and legal frameworks, including ethical decision-making models to which ETHAI belongs, will be necessary to proceed with the CE marking, mandatory for placing any AI-based systems in the EU market. At the same time, standardisation committees ISO/IEC JTC 1/ SC 42 "Artificial Intelligence" and CEN-CLC/JTC 21 focus on producing standards that address market (e.g. interoperability) and societal needs (e.g. ethics assessment), as well as underpinning EU legislation, policies, principles and values.

The ETHAI model defined in MES-CoBraD is moving towards a refinement and enhancement process, based on assessment that will be performed among four pilot use cases.

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 965422.

Links:

[L1] https://mes-cobrad.eu/
[L2] https://cyberethicslab.com

References:

[1] "History of flu (influenza): Outbreaks and vaccine timeline." Mayo Foundation for Medical Education and Research (MFMER). https://kwz.me/hxE (accessed 2023).
[2] "Health and population." United Nations. https://sdgs.un.org/topics/health-and-population (accessed 2023).
[3] EPRS/STOA, "Privacy and security aspects of 5G technology," 2022. Available at: https://doi.org/10.2861/255532

Please contact:

Luigi Briguglio, CyberEthics Lab., Italy
l.briguglio@cyberethicslab.com

Current Challenges and Future Research Directions in Multimodal Explainable Artificial Intelligence

by Nikolaos Rodis, Christos Sardianos and Georgios Th. Papadopoulos (Harokopio University of Athens)

As Artificial Intelligence (AI) continues to advance and find applications in various domains, the need for explainable AI becomes crucial. In the field of multimodal explainable AI (MXAI), which deals with multiple types of data, challenges arise in defining terminology, utilizing attention mechanisms, generalizing methods, extending explanations to more modalities, estimating causal explanations, and removing bias. Addressing these challenges is essential for improving transparency and trustworthiness in critical domains like healthcare.

Despite the outstanding advances in AI and its widespread adoption in several application domains, there are still significant challenges that need to be addressed regarding the explanation of how decisions are reached to the end-user. The latter need becomes more complex and demanding when multiple types of data are involved in the AI-based generated decisions; hence, leading to the emergence of the so-called multimodal explainable MXAI field. The above challenges become even more imperative for some critical domains, for example, medical applications (where human lives are involved).

The capabilities of the medical sector have recently been tremendously improved, largely due to the introduction of multiple AI-boosted applications for improving, for example, decision-making, drug development and disease prevention. However, explaining the obtained results is not always easy, with the involvement of multiple data modalities acting as an extra obstacle, as graphically illustrated in Figure 1. Within the context of the EC-funded project ONELAB [L1], diverse and heterogeneous information sources will be used for addressing the issue of biomarker detection in breath data, for example, gas chromatography–ion mobility spectrometry (GC-IMS), and gas chromatography–mass spectrometry (GC-MS). Towards this goal, MXAI approaches are employed to provide the required explanations of the obtained AI-based predictions. The challenges of this problem are numerous, and brief descriptions of some of them are presented below.

Convergence to Formal and Widely Accepted Definitions/Terminology

Despite numerous research works being recently introduced in the MXAI field, little-to-no formality has been introduced concerning the adopted definitions and terminology. In particular, many researchers make use of ad-hoc descriptions to delineate their research activities, while they often define "explainability" and "interpretability" in various ways. As a result, no exact and widely accepted terminology is present in the field. Defining what an explanation is and how its efficiency can be measured (based on both qualitative and quantitative norms and experimental frameworks), apart from enhancing formalisation aspects in the field, will also significantly facilitate the comparative evaluation of the numerous proposed MXAI methods. The latter will also greatly assist in addressing current controversies, like assigning different terms to similar methods or associating similar names with fundamentally different (algorithmic) concepts.

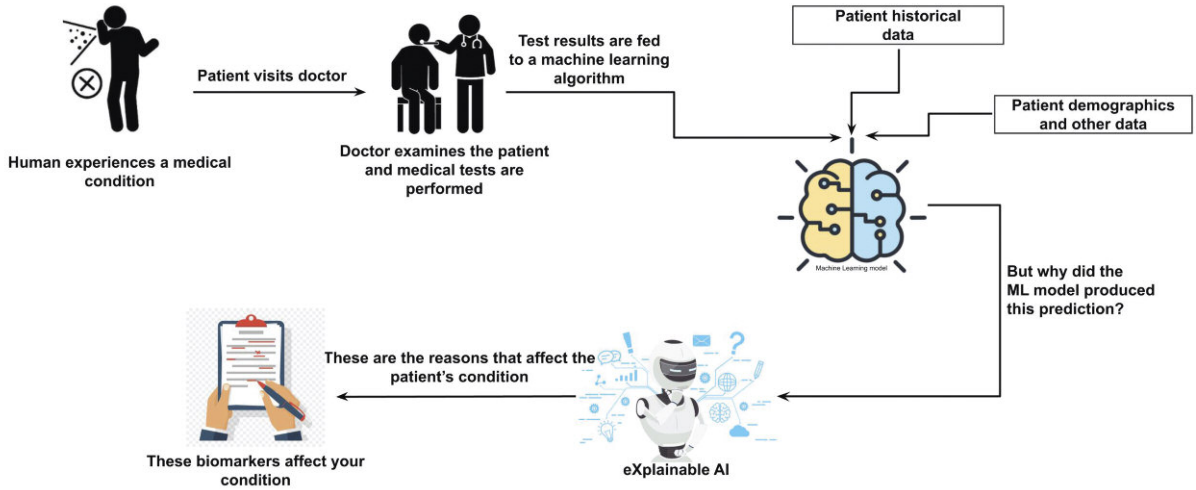


Figure 1: A visual representation of a patient-centric data flow: enabling AI-driven biomarker analysis and explainable insights for doctors

Usage of Attention Mechanisms in Explanation Schemes

Attention schemes, apart from being used in numerous data-analysis tasks, have also been utilised for generating explanations of corresponding prediction models, typically in the form of visualisation methods (indicating text segments or image areas where the primary AI prediction model focuses) or estimating feature-importance metrics. However, several concerns and controversies have emerged, fundamentally raising doubts regarding the suitability of attention mechanisms to produce actual explanations. In particular, experimental studies show

that attention distributions between learned attention weights and gradient-based feature relevance methods are not highly correlated for similar predictions [1]; hence, conventional attention explanations cannot be considered equivalent to others. However, contradictory experimental results in more recent works move to the opposite direction, that is, the usage of attention schemes for explanation generation is not always applicable, but it depends on the actual definition of explanation that is adopted in the particular application at hand. In this context, more detailed and in-depth studies need to be conducted, in order to shed more light on whether and under which exact conditions certain attention schemes can be used for providing meaningful explanations, as well as how such methods relate to other non-attention-based MXAI approaches.

Generalisation Ability of MXAI Methods

The wide majority of the available methods have only been designed for specific AI model architectures (regarding the primary prediction task) and in many cases they are constrained to specific analysis tasks [2]. For example, there is a significant number of methods that have been designed for the particular visual question answering (VQA) task; however, such approaches have not been evaluated in other vision-language applications. Naturally, it can be well admitted that introducing model-specific explanation schemes is very restrictive and expensive. Robustly extending existing methods to other tasks and architectures would significantly reduce research and development efforts.

Extension of MXAI Schemes to More than Two Output Modalities

The wide majority of MXAI methods focus on producing unimodal or bimodal explanations. However, extending explanation representations to higher dimensionality multimodal feature spaces (i.e. feature spaces that are composed of more than two modalities) would inevitably further increase the expressiveness and accuracy of the produced interpretations.

Estimation of Causal Explanations

So far, no significant attention has been given to the causality perspective of explanations, while causal relationships are the main type of relationships that humans inherently perceive. In this respect, causal explanations can enable the understanding of how one event can lead to another and, hence, to develop a deeper understanding of the world. On the other hand, identifying the factors that cause an event to occur can also facilitate how the event might unfold and/or how it should be encountered [3]. Therefore, apart from identifying which features are important for a given model, how predictions are affected from modification in the feature values is important to understand the model's reasoning process itself. In the context of the multimodal setting, causality needs to be examined in terms of how each individual modality and the corresponding features affect the prediction outcome (and not simply identifying which features are important).

Removing Bias in Textual Explanations

The main paradigm being followed for estimating textual explanations consists of collecting natural language rationales from humans (for a given dataset) and subsequently developing/training an explanation module with these descriptions as ground truth. However, human textual annotations (especially when it comes to long textual justifications) typically contain (contradictory) biases that are related to the particular background and temperament of each involved individual. To this end, developing routines for identifying/removing bias and resolving conflicting annotation cases would also significantly improve the quality of the respective generated textual annotations.

To conclude, further research needs to be conducted regarding the above-mentioned problems in order to make AI models more transparent, trustworthy and to boost their utilisation in critical domains (e.g., health care, self-driving cars, etc.).

The research leading to the results of this paper has received funding from the European Union's Horizon Europe research and development programme under grant agreement No 101073924 (ONELAB). The authors would also like to thank Prof. Iraklis Varlamis for his valuable guidance and comments on formulating the above challenges and open issues.

Link:

[L1] https://onelab-project.eu

References:

[1] S. Jain, B. C. Wallace, "Attention is not explanation," in North American Chapter of the Association for Computational Linguistics, vol.1, Jun. 2019, pp. 3543–3556.

[2] G. Joshi, R. Walambe, K Kotecha, "A review on explainability in multimodal deep neural nets," IEEE Access, vol. 9, 2021, pp. 59800–59821.

[3] S., Waddah, C. Omlin, "Explainable ai (xai): a systematic meta-survey of current challenges and future opportunities," Knowledge-Based Systems, vol. 263, 2023, doi: 110273.

Please contact:

Nikolaos Rodis, Harokopio University of Athens, Greece
rodisnick@hua.gr

Christos Sardianos, Harokopio University of Athens, Greece
sardianos@hua.gr

Predictive Model for Functional Outcome after Orthopaedic Surgery Using Machine Learning Methods

Alexandre Lädermann (Hôpital de La Tour, Meyrin, Switzerland), Philippe Collin (American Hospital of Paris, France) and Patrick J. Denard (Oregon Shoulder Institute, Medford, Oregon, USA)

Surgery is said to be indicated when conservative treatment fails. Previous studies reported that around 20% of patients do not improve sufficiently after surgery, inducing frustration, high societal costs, and an overload of healthcare systems. The consortium members investigated the efficacy of machine learning methods in detecting outcomes. They achieved a promising model with a recall of 32% of the cases that were inappropriate candidates for an operation.

Healthcare systems face several challenges linked to the ageing of the population and an increase in the prevalence of chronic conditions. Among them, debilitating musculoskeletal pathologies are widespread, resulting in severe restrictions on daily activities and work capabilities, leading to a dramatic increase in the number of patients admitted to hospitals for surgery during their last decades. These musculoskeletal pathologies constitute a high societal cost and an occupational burden for workers; however, there is little overall evidence behind surgical indications. Consequently, around one-tenth to one-fifth of them will be not treated in the appropriate setting according to their respective medical condition in a healthcare continuum.[1] Some patients could have avoided surgeries and have been sent directly from inpatient to outpatient care services such as physiotherapy, improving care pathways instead of compromising use of resources.

To solve these problems, our consortium developed an innovative decision-making tool that could improve clinical practice guidelines and help communicate the expected result from a proposed surgical treatment as an essential component of informed consent. Machine learning (ML) is a field that focuses on the learning aspect of artificial intelligence (AI) by developing algorithms that best represent a set of data.[2] Within our consortium, existing collaborating partners are used to apply value-based health care (VBHC) strategy and skills.[3] Evaluating the value of health care is of paramount importance to keep improving patients' quality of life and optimising associated costs. This study aimed to compare ML algorithms and determine the accuracy of AI in predicting clinical outcomes after rotator cuff repair. The hypothesis was that preoperative clinical and intraoperative data alone would be insufficient to provide proper guidelines.

We analysed prospectively collected data from patients undergoing rotator cuff repair between March 2013 and July 2020. In total, 9,030 cases were enrolled, leading to a refined selection of cases: 4,683 subjects, which were divided into the training (80%) and testing (20%) sets. Different ML models were trained and tested using the 236 valid pre-treatment features. The models were also tested on a reduced set of only ten features, identified through Shapley Additive Explanations (SHAP) (Figure 1). The Single Assessment Numeric Evaluation (SANE) score at one year was the output variable. Minimal specificity was set at 95%.

The performance of the XGBShap10 model revealed a specificity of 0.951 (0.934–0.965), a precision of 0.587 (0.485–0.685), a recall of 0.32 (0.252–0.392), an accuracy of 0.838 (0.814–0.861), an F1 score of 0.413 (0.338–0.487) and an area under the receiver operating characteristic curve (AUC) of 0.667 (0.618–0.715).

With a mean accuracy of 84% and a specificity set above 95%, our pilot study showed that the model is not a simple heuristic; it could be integrated into existing healthcare information systems to help clinicians develop better and more reasonable treatment programmes, more adequately inform patients about expected results (empowerment), and save, at a European

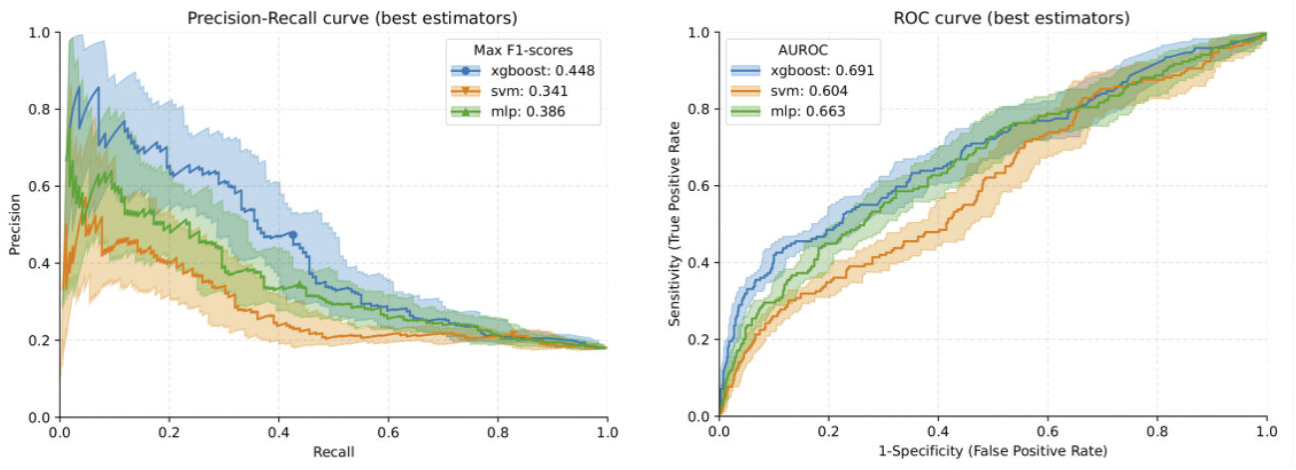


Figure 1: Plots for the precision-recall curve (left) and the area under the ROC curve (right) for the three machine learning models: XGBoost, multi-layer perceptron (MLP) and support vector machine (SVM). Mean ROC is represented in the AUROC as a plot line, with the standard deviation of the cross-validation shown as a shade.

level, billions per year. We now aim to improve the accuracy and fairness of the tool, obtain certification, develop better guidelines and transfer such platform technology to other disciplines, such as knee (anterior cruciate ligament) and foot (Achilles tendon) surgeries.

The authors would like to acknowledge the consortium partners Med4Cast (Martigny, Switzerland), Idiap Research Institute (Martigny, Switzerland), the insurance Group Mutuel (Martigny, Switzerland), the IDE4 foundation (Geneva, Switzerland), and the Image Analysis research unit of the Université Libre de Bruxelles (LISA-IA, Belgium), Katalysen (Stockholm, Sweden), SCIPROM (Lausanne, Switzerland), and Nexialist (La Ciotat, France) for their valuable feedback in the design, implementation, visual-data acquisition, certification and annotation involved in this project.

This project has received funding from the Fondation de Bienfaisance Pierre & Andrée Haas (Geneva, Switzerland), the insurance Groupe Mutuel (Martigny, Switzerland), and FORE (Foundation for Research and Teaching in Orthopedics, Sports Medicine, Trauma, and Imaging in the Musculoskeletal System), Grant #2023-13.

References:

[1] P. Collin, et al., "Prospective evaluation of clinical and radiologic factors predicting return to activity within 6 months after arthroscopic rotator cuff repair," Journal of Shoulder Elbow Surgery, vol. 24, no. 3, pp. 439–45, Mar. 2015, doi: 10.1016/j.jse.2014.08.014.

[2] L. J. H. Allaart et al., "Developing a machine learning algorithm to predict probability of retear and functional outcomes in patients undergoing rotator cuff repair surgery: protocol for a retrospective, multicentre study," BMJ Open, vol. 13, no. 2, p. e063673, Feb. 10 2023, doi: 10.1136/bmjopen-2022-063673.

[3] A. Lädermann, et al., "Measuring patient value after total shoulder arthroplasty," Journal of Clinical Medicine, vol. 10, no. 23, Dec. 4 2021, doi: 10.3390/jcm10235700.

Please contact:

Alexandre Lädermann, Division of Orthopaedics and Trauma Surgery, Hôpital de La Tour, Meyrin, Switzerland
alexandre.laedermann@gmail.com

Unleashing the Power of Artificial Intelligence for Personalised Drug Design

by Michaela Areti Zervou, Effrosyni Doutsi, Panagiotis Tsakalides (University of Crete and ICS-FORTH)

Precision medicine holds the promise of personalised treatment based on an individual's genetic makeup. However, the lengthy and costly process of drug discovery hinders progress. Can artificial intelligence (AI), specifically generative models offer a solution? Is it possible to efficiently select the most promising drug candidates for validation? Our research focuses on developing robust tools that streamline the validation process of drug design – saving time and resources.

Precision medicine [1] represents a new era in health care, aiming to provide tailored treatments based on an individual's unique genetic composition. By analysing an individual's DNA, healthcare providers can make decisions about the most suitable treatment. However, significant challenges in developing new drugs, such as the time consumption and the considerable cost of the design process, often involves years of trial and error with millions of compounds being tested.

A crucial task for efficient drug design is accurately predicting the position of the target protein within the cell. This knowledge empowers the design of drugs that specifically target an intended protein, disrupt or enhance its function while minimising the risk of unintended effects on healthy cells. This holds the key to developing highly targeted drug candidates that can bind with exceptional precision, maximising therapeutic benefits, while reducing potential adverse reactions.

The power of artificial intelligence (AI) offers promising opportunities to address the challenges in predicting protein position and designing drugs with enhanced specificity. But how can we leverage the power of AI to predict protein position and design drugs that specifically target intended proteins while minimising off-target effects? The answer lies in the realm of generative models [2]. These models are a fascinating blend of mathematics, statistics and computer engineering. They are trained on large databases of existing molecules trying to identify and learn the underlying patterns and relationships between molecular structures. Based on the acquired knowledge, they allow generation of a vast number of novel molecules that adhere to the learned patterns while introducing variations. This enables researchers to explore a wide range of potential drug candidates, but how can we ensure the reliability and effectiveness of all these generated molecules? Moreover, is it possible to accelerate the validation process in the laboratory?

To address these challenges, our research funded by the Hellenic Foundation for Research and Innovation (HFRI) under the Ph.D. Fellowship grant no. 5647 [L1], focuses on developing robust classification tools to effectively learn and classify the position of the proteins in the cell. As shown in

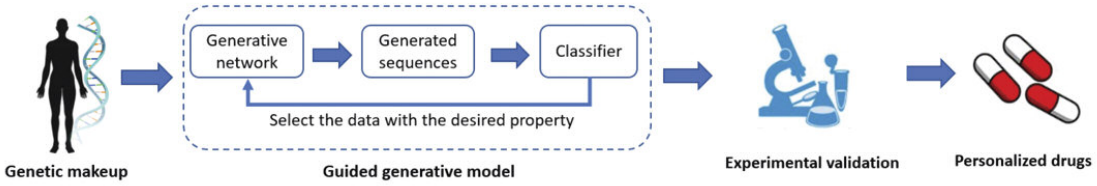


Figure 1: Personalised drug-design architecture pipeline.

Figure 1, by integrating these tools into the de novo design pipeline, we can guide the process of designing to produce novel molecules with a desired position of the target protein in the cell. This approach enables researchers to prioritise and select the most promising candidates for experimental validation, thereby streamlining the validation process. Ultimately, this reduces the number of potential designs that require synthesis and testing, leading to significant time and resource savings.

Our research team has made significant advancements in the development of novel classification tools that surpass the current state-of-the-art [3]. Leveraging innovative algorithms and machine learning techniques, we have created highly accurate tools that can effectively identify the position of the target protein in the cell.

Furthermore, to bridge the gap between research and practical applications, we have established collaborations with renowned biologists and experts in the field. These collaborations enable us to validate the effectiveness of our classification tools in real-world scenarios, leveraging their domain knowledge and expertise. By aligning our research with industry requirements and working closely with experienced biologists, we aim to develop practical solutions that have the potential to revolutionise the drug-discovery process.

In conclusion, precision medicine holds immense potential for personalised health care. However, the lengthy and costly drug-discovery process poses significant challenges. The power of artificial intelligence and generative models presents an opportunity for a breakthrough in this domain. Efficiently selecting the most promising drug candidates for validation is crucial to saving time and resources. Through our research and collaborations with experts in the field, including renowned biologists, we are working towards a more efficient and effective drug-design process. By harnessing the capabilities of artificial intelligence and leveraging generative models, we aim to revolutionise the development of personalised treatments. Ultimately, our efforts bring us one step closer to realising the transformative impact of personalised medicine on patients worldwide.

Links:

[L1] https://www.elidek.gr/en/call/3rd-call-for-h-f-r-i-scholarships-for-phd-candidates/

References:

[1] M.R. Kosorok and E. B. Laber, "Precision medicine," Annual Review of Statistics and its Application, vol. 6, pp. 263–286, 2019.

[2] X. Pan and T. Kortemme, "Recent advances in de novo protein design: principles, methods, and applications," Journal of Biological Chemistry, vol. 296, 2021.

[3] M. A. Zervou, E. Doutsi and P. Tsakalides, "Efficient protein structural class prediction via chaos game representation and recurrent neural networks," 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094877.

Please contact:
Michaela Areti Zervou
University of Crete and ICS-FORTH, Greece
zervou@ics.forth.gr

Effrosyni Doutsi, ICS-FORTH, Greece
edoutsi@ics.forth.gr

Merging Explainable AI into Automotive Software Development

by Danilo Brajovic and Marco F. Huber (Fraunhofer Institute for Manufacturing Engineering and Automation IPA)

The debate about reliable, transparent and thus, explainable AI applications is in full swing. Despite that, there is a lack of experience in how to integrate AI-specific safety aspects into standard software development. In the veoPipe research project, Fraunhofer IPA, Huber Automotive, and ROI-EFESO Management Consulting work on a joint approach to integrate these AI-specific aspects into an automotive-development process. In this article, we share details on one component of this framework – reporting the AI development.

The emerging regulation of Artificial Intelligence (AI) by the EU AI Act requires developers and companies to adjust their processes in order to be able to develop trustworthy AI systems. However, there is a lack of experience in how this integration can be accomplished, and there are still several technical challenges to be solved. Fundamentally, what makes development of trustworthy AI systems more difficult than standard software is the fast-paced development and the black-box character of many methods. This means that even experts are often unable to adequately understand how exactly an artificial neural network or any other machine learning model arrived at a particular result. This is unfavourable in several respects. It reduces user confidence in the AI application. AI experts or

developers have little opportunity to specifically improve the application if even they do not fully understand the application. Finally, there are legal challenges around liability in the case of failures if the reason for the failure cannot be clarified. These challenges are addressed by the research field of explainable AI (XAI) [1].

On the industry side, there are several pieces of work to establish guidelines for developing trustworthy AI systems. Especially among US companies, reporting datasets and machine learning models has become an established measure [2,3]. However, most of these works only cover a single stage of AI development, lack technical details especially regarding XAI, or are not aligned with regulation. Our contribution is a structured reporting framework based on a set of cards that extends prior work and goes along four major development steps. Our cards not only provide a structured information overview on the corresponding development step, but also refer to additional material such as regulatory standards, toolboxes or scientific publications. With these interim results, developers should already be very well positioned when legal requirements for the use of AI applications come into force. The four developments steps covered in our approach are:

1. Use Case Definition: This card describes the intended application in more detail and identifies potential risks, e.g. under the EU AI Act.
2. Data: This card deals with the documentation and collection of data, labelling, its provision and pre-processing.
3. Model Development: This card covers topics such as interpretability and explainability, model selection, training, evaluation, and testing.
4. Model Operation: Once the model is practically in use, issues such as the concept of operation, model autonomy, monitoring, or adversarial attacks and data protection become relevant.

We evaluated our proposal on three use-cases within the automotive domain. Further, in the course of our project, we held interviews with several certification bodies and companies in order to understand their approach and requirements to safeguard AI systems and, in particular, to use XAI for this purpose. Our industry partners revealed that they had experimented with XAI methods but found them to be too unreliable for practical use. If any, they can help to find model errors, but they provide no help in safeguarding the application. On the other side, certification bodies mentioned that it is hard to define specific requirements due to the rapidly evolving field, but the best thing companies can do is to show that they recognised the problem, for example, by applying methods such as Local Interpretable Model-agnostic Explanations (LIME) or Shapley Additive Explanation (SHAP) .

Currently, a full paper describing the approach is being prepared and will be submitted soon. We are also integrating it into a classical automotive V-model and plan a follow-up study with certification bodies to better understand their approach to XAI.

References:

[1] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," Journal of Artificial Intelligence Research (JAIR), vol.70, pp. 245–317, 2021.

[2] Mitchell et al., " Model cards for model reporting", in Proc. Conf. on Fairness, Accountability, and Transparency (FAT), 2019.

[3] M. Pushkarna et al., "Data cards: purposeful and transparent dataset documentation for responsible AI", in Conf. on Fairness, Accountability, and Transparency (FAccT), 2022.

Please contact:

Danilo Brajovic, Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Germany
danilo.brajovic@ipa.fraunhofer.de

The project team with the test vehicle. Within the project, the developed approach is evaluated on two use-cases. One of these use-cases involves pedestrian tracking on a vehicle.

An Explainable Deep Ensemble Framework for Intelligent Ticket Management

by Gianluigi Folino, Massimo Guarascio, Luigi Pontieri and Paolo Zicari (CNR-ICAR)

Pushing intelligence and integrating explainable tools in the new generation of ticket-management systems is crucial for supporting customer-support activities. To this aim, we defined a comprehensive ticket-classification framework, which integrates deep ensemble methods and AI-based interpretation techniques to help both the operator identify misclassification errors and the analyst improve the model. Tests on real data demonstrate the quality of the predictions returned by the framework and the practical value of their associated explanations.

Nowadays, ticket-management systems (TMS) are widely used to improve the organisation, efficiency and effectiveness of customer support, with relevant impacts on costs and revenues, customer retention and public brand image. In particular, equipping these systems with intelligent tools able to provide reliable classifications and speed up the assignment process represents a challenging task for both industrial and academic research, as it requires coping with several issues (e.g. data scarcity, noise and skewness on data and processing natural language).

Tickets opened by a customer request through different channels (e.g. phone calls, emails, web forms, live chats and recently also social media like Facebook and Twitter) can be differently routed based on their properties, for example, the urgency, impact, specific area of interest or resource allocation scheme. Natural language processing methods and machine learning techniques have been widely used in the literature to develop automatic ticket classification approaches and boost customer-support systems' capacities.

In this respect, deep learning (DL) is effectively and efficiently used to process text data [1]. Despite the great potential of DL-based text classifiers, the performances of the current solutions can be affected by different challenging issues frequently occurring in real-life applications. First, large corpora of labelled data are usually necessary to adequately train a deep model (the deeper and more complex the model, the more example data are needed). Moreover, configuring the topology and hyper-parameters of a deep neural network (DNN) architecture is a difficult task that entails long and careful design and tuning activities to make the DNN perform well. Finally, training data typically exhibit an unbalanced distribution, that is, some classes are more frequent than others (e.g. in the TMS scenario, tickets with high urgency are less frequent than the ones with lower urgency). These issues entail an increased risk of learning DNN-based classifiers that overfit the training data and rely on non-general, biased and unreliable classification patterns hinging on spurious features. Moreover, the black-box nature of a DNN model does not allow an easy understanding of which features of a data instance drove the model to its classification decision.

To cope with these issues, we have defined a ticket-classification framework based on ensemble DNN classifiers, leveraging different types of neural architectures (LSTM, CNN, GRU and transformers) as base classifiers that promote diversity, expressiveness, and robustness to over-fitting and class-imbalance risks. The framework introduces two novel ensemble

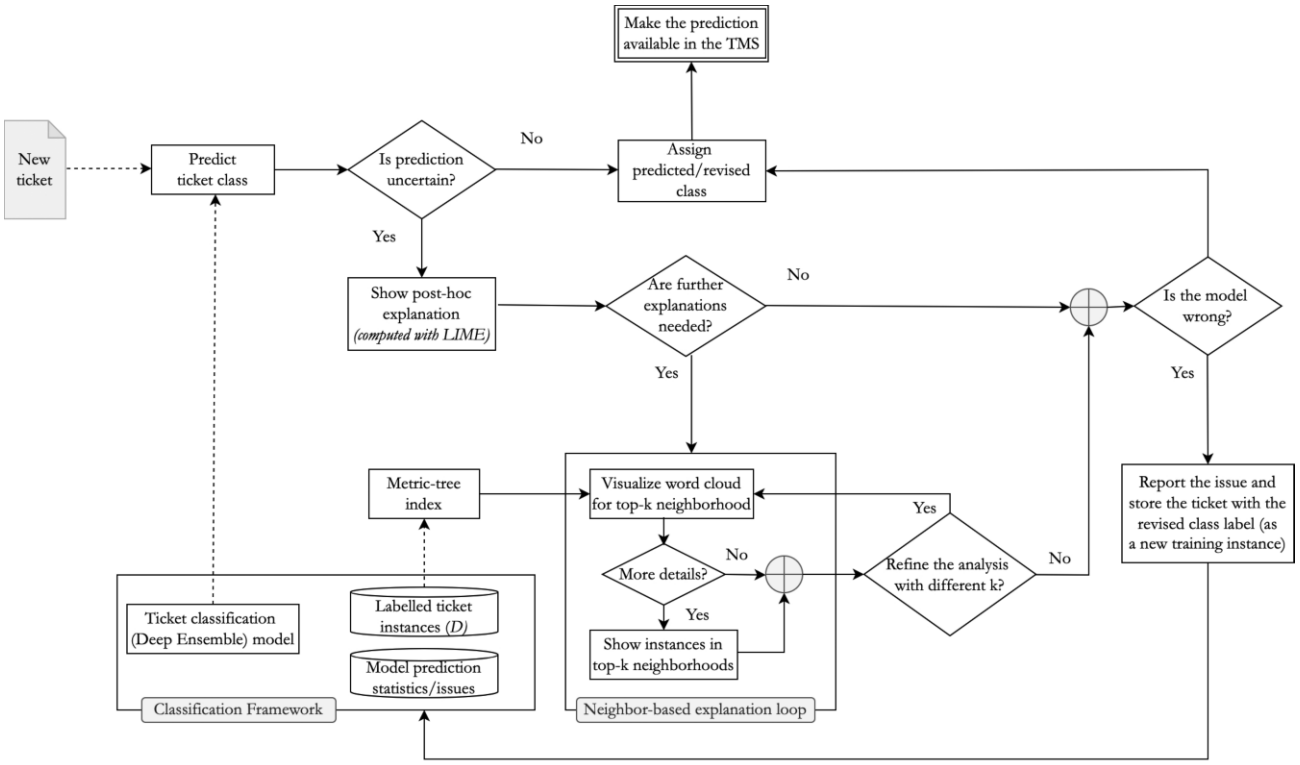


Figure 1: The Human-in-the-loop scheme of the proposed intelligent ticket classification framework.

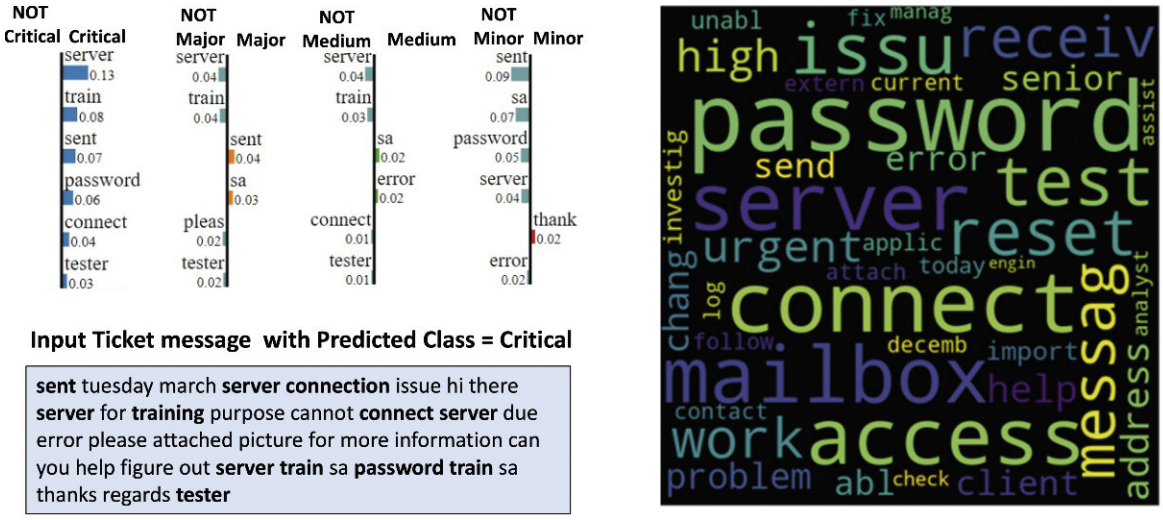


Figure 2: The explanation output (left: LIME-based explanation; right: word cloud).

combination strategies based on stacking and mixture-of-expert (MoE) architectures, both leveraging an ad hoc sub-net, named High-LevelFeature Extractor (HLFE), which allows for extracting compressed text representations (latent space). The explanation module described in Figure 1 is integrated within the framework to support a continuous, human-in-the-loop scheme for discovering, using, validating and improving the ensemble DL model.

In more detail, two kinds of explanation artefacts are provided. Based on the post-hoc explanation algorithm LIME, the former extracts the subset of terms that likely drove the model predictions. The latter yields per-class word-cloud representations, obtained by computing the distributions of words in neighbours of the test instance that belong to each class.

Tests assessed the effectiveness and usefulness of the proposed framework over two real-life and unbalanced ticket datasets furnished by customer-support companies. The first dataset is related to a phone-company customer-care ticket collection in the form of SMS messages and short Facebook chat texts. The second dataset is a publicly available customer-support ticket dataset containing about 50 thousand tickets submitted via email by the customers of the Endava company to the helpdesk. Comparison tests demonstrated evident improvements in terms of the F1, AUC and G-measure metrics, obtained by using the novel deep ensemble approach, both in comparison with the baseline DNNs and with other state-of-the-art machine learning and ensemble-based methods (i.e. Gradient Boosting, Random Forest). For all the competitors in the comparison, the implementations available in the popular machine learning library scikit-learn were adopted, with settings chosen after performing a grid-search procedure.

The proposed explanation workflow was validated through several tests furnishing the LIME representation and the word cloud of the k nearest neighbour tickets found by using the cosine similarity in the latent-space representation of the tickets. An explanation obtained with our LIME-based procedure (on the left) and with the word cloud (on the right) is shown in Figure 2 for a ticket classified in the critical (really urgent) class. It is evident that the terms that LIME deems most influential for this prediction decision (highlighted in blue in the ticket message) confirm that the model focuses on concepts (namely: server, connection, password, tester, sent) that are really relevant for deciding the class of the ticket. The relevance of these terms was confirmed by the word cloud derived from the $k = 40$ nearest neighbours.

Moreover, the *k*-nearest neighbour-based explanation supports the customer operators in explaining classification decisions by showing similar labelled examples. Thus, all these explanation artefacts support the model's prediction for the considered ticket. The explanation quality tests demonstrated the practical added value given by the proposed explanation artefacts to the operators in improving the classification model and the customer services, thus providing a toolset of interpretation methods helping both the operator in recognising misclassification errors and the analyst in improving and fine-tuning the model.

This work is a result of research activities conducted in the context of the H2020 project ''HumanE-AI-Net'' [L1], funded by the European Commission (grant no. 952026).

Link:

[L1] https://cordis.europa.eu/project/id/952026/it

References:

[1] S. Minaee et al., "Deep Learning-based Text Classification: A Comprehensive Review", ACM Computing Surveys 54(3): 62:1–62:40, 2022

[2] P. Zicari et al., "Discovering accurate deep learning based predictive models for automatic customer support ticket classification," in Proc. ACM SAC, 2021, pp. 1098–1101. Available: https://doi.org/10.1145/3412841.3442109

[3] P. Zicari et al., "Combining deep ensemble learning and explanation for intelligent ticket Management,"Expert Systems with Applications, vol. 206, 2022. Available: https://doi.org/10.1016/j.eswa.2022.117815

Please contact:

Gianluigi Folino, CNR-ICAR, Italy
gianluigi.folino@icar.cnr.it

Decoding the Unknown: Unveiling Industrial Time Series Classification with Counterfactuals

by Anahid Jalali (AIT), Andreas Rauber (TUWien), Jasmin Lampert (AIT)

Deep learning models for time series prediction have become popular with the rise of IoT and sensor data availability. However, their lack of explainability hampers their use in critical industrial applications. While existing model-agnostic approaches like LIME and SHAP have been used in time series classification applications, it is worth mentioning that they may have limitations in their suitability. For example, the random sampling process used by LIME leads to unstable explanations. We propose a counterfactual explanation approach for interpretable insights into time series predictions to address this issue. We choose an industrial use case, determining machine health, and employ k-means clustering and Dynamic Time Warping (DTW) to handle the temporal dimension. DTW compares and aligns two time series by discovering the optimal path of alignment that minimizes disparities in their temporal patterns. We explain the model's decisions using local surrogate decision trees, analysing feature importance and decision cuts.

For this purpose, we focus on a time series classification task that determines whether a machine is healthy or unhealthy. We introduce an automatic example-based approach to extract factual and counterfactual samples from clustered input data, enabling justification of model classification results. To achieve this, we employ the k-means algorithm to partition the time series into clusters, minimising clustering error by considering the sum of squared distances from each data point to its cluster centre. Unlike spatial distance-based approaches, we utilise Dynamic Time Warping (DTW) as a similarity measure to account for the temporal dimension inherent in time series data.

To explain the model's decisions, we identify each cluster's most important time series features, including frequencies, amplitude, pitch, mean and standard deviation. We further employ local surrogate decision trees (DTs) as interpretable models to explain the black-box decisions. DTs are well-known for their interpretability and require fewer resources than other methods. We elucidate the temporal changes and their contributions to the prediction results by analysing these DTs' feature importance and decision cuts. The tree's feature importance is used to identify the top influential parameters, and the decision cuts help extract classification decision boundaries.

We tested our proposed approach on the Commercial Modular Aero-Propulsion System Simulation (CMAPS) dataset [L1], a widely cited dataset used for Prognosis Health Management tasks. The dataset consists of 218 engines that start in a healthy state and experience artificially injected faults until breakdown. We have experimented with different cluster sizes, evaluating them using the silhouette metric and determining the optimal number of clusters to create smaller neighbourhoods.

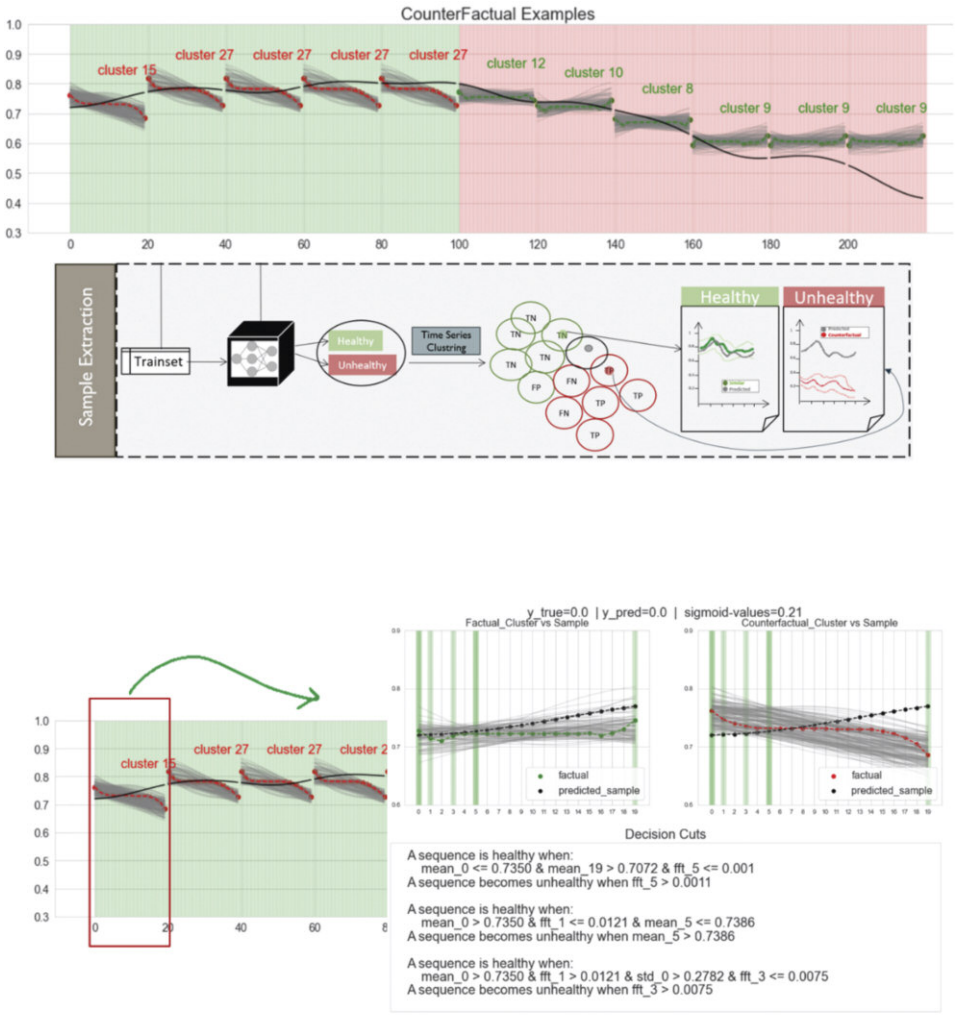


Figure 1: Illustration of the extracted counterfactual samples for one engine, in which the surrogate model had 100% accuracy in predicting its health state. The background colour indicates the model prediction for each flight sequence: green for healthy and red for unhealthy. This plot shows the dissimilarities of the counterfactual to the sequence, e.g. when the predicted unhealthy cluster have different behaviour in both spatial and time dimensions compared to the predicted sequences [4].

We calculate the closest clusters representing healthy and unhealthy conditions for each test sequence. We then utilise a trained LSTM model to predict the class of the test sample and, based on the prediction, assign the factual and counterfactual clusters accordingly. We repeat this process for all sequences of one engine.

Additionally, we investigate the contribution of time series characteristics to the classification output. For this, we extract time-domain features such as mean, standard deviation, minimum and maximum from each cluster and train local surrogate DTs on these clustered features. The feature importance analysis reveals that the mean values at the signal's first and last time steps are the most influential features in predicting the healthy or unhealthy classes. Other significant features include the lower frequency of the time series and the standard deviation of the second half of the sequence's time steps. By extracting rules from the DTs, the authors gain insights into parameter changes and their impact on decision-making.

In Figure 1, we present visual representations of factual and counterfactual examples. A factual example includes: i) the test sequence sample, ii) the centre of the closest cluster with the same class, and iii) all the samples in the cluster. Similarly, a counterfactual example includes: i) the test sequence sample, ii) the centre of the closest cluster with the opposite class and iii) all the samples in the cluster. These visualisations help illustrate the dissimilarities between the clusters and provide a further understanding of the model's decisions.

As our research progresses, we are expanding our approach to encompass time series forecasting. Additionally, we strongly advocate incorporating expert feedback into the explanations, as it can enhance both the model's performance and the overall quality of the explanations.

Links:
[L1] https://www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository

References:
[1] T. Sivill and P. Flach, "LIMESegment: meaningful, realistic time series explanations," in Int. Conf. on Artificial Intelligence and Statistics, 2022, pp. 3418–3433.
[2] M. Guillemé, et al., "Agnostic local explanation for time series classification", in 2019 IEEE 31st Int. Conf. on Tools with Artificial Intelligence (ICTAI), Nov. 2019, pp. 432–439, 2019
[3] A. Theissler, et al., "Explainable AI for time series classification: a review, taxonomy and research directions," IEEE Access, 2022.
[4] A. Jalali, et al., "Explaining binary time series classification with counterfactuals in an industrial use case," presented at ACM CHI Workshop on Human-Centered Perspectives in Explainable AI, 2022.

Please contact:
Anahid Jalali, AIT Austrian Institute of Technology, Austria
anahid.jalali@ait.ac.at

ChatGPT Responses Validation through Knowledge Graphs

by Michalis Mountantonakis and Yannis Tzitzikas (FORTH-ICS and University of Crete)

The novel artificial intelligence ChatGPT chatbot offers detailed responses across many domains of knowledge; however, quite often it returns erroneous facts even for popular persons, events and places. To tackle this problem, we present GPT•LODS, a novel prototype that annotates and validates ChatGPT responses by leveraging one or more RDF knowledge graphs.

There is a recent trend for using the novel artificial intelligence ChatGPT chatbot, which is an innovative application of large language models (LLMs) that provides detailed and articulate responses across many domains of knowledge. However, in many cases, it returns plausible-sounding but incorrect or inaccurate responses, it does not provide justifications, and its current version has "limited knowledge of world and events after 2021". On the other hand, there is a high proliferation of knowledge graphs (KGs) that are modeled using the Resource Description Framework (RDF) model over any real domain. These KGs offer high-quality structured data by recording their provenance, whereas most of the popular RDF KGs are updated at least periodically. Therefore, the key notion is how to enable the combination of ChatGPT and RDF KGs, for making it feasible to enrich and validate any ChatGPT response, and this is quite challenging since it requires access to numerous RDF KGs, sources and resources in general.

The Information Systems Laboratory of the Institute of Computer Science of FORTH designs and develops innovative algorithms and tools for enabling the combination of ChatGPT and RDF KGs. We call the corresponding services "GPT•LODS" [1], [L1] (the name of GPT•LODS stems from the mathematical notation for function composition). The key idea, illustrated in Figure 1, is to send a question to ChatGPT, which has been trained by using data from web sources (such as Wikipedia, books and news articles), and then to enrich its response by using hundreds of RDF KGs through LODsyndesis [2], [L2], which aggregates data from hundreds of RDF KGs from several domains, containing in total more than 2 billion triples. The current version of GPT•LODS (accessible online in [L1], also offering tutorial videos), provides two different types of services: (i) an annotation and enrichment service for enabling the identification, linking and enrichment of the entities of a ChatGPT response, and (ii) a fact-checking service for validating the facts of a ChatGPT response, and accompanying them with provenance information. Below we describe these services through an example where we ask ChatGPT "Who was the scorer of the UEFA Euro 2004 Final", as shown in Figure 2.

The annotation and enrichment service retrieves a ChatGPT response and offers real-time annotation, linking and enrichment of its entities based on hundreds of RDF KGs, by using natural language processing tools (specifically named entity

recognition and linking tools). In the example of Figure 2, the system managed to identify and link the entities of the response, in particular, it found more information (links, images, entity type, datasets and facts) for each of the entities of the response, for example, for the entity "Greece" it found 261 URIs and 87 thousand facts from 40 RDF KGs. By clicking on one of those links, one can browse all this information (see the lower part of Figure 2), for example, see on the left side all the URIs for the entity "UEFA Euro 2004 Final" and on the right side the RDF datasets (or KGs) including information about "Greece".

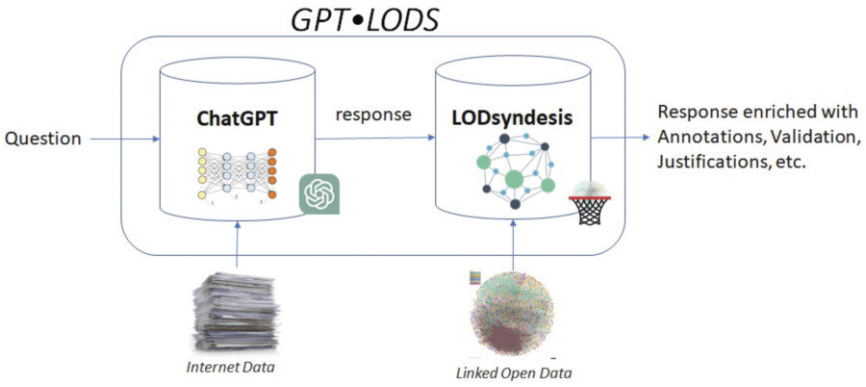


Figure 1: The key notion of combining ChatGPT with RDF knowledge graphs.

However, the key problem of the ChatGPT response of Figure 2 is that it contains some erroneous facts: the scorer of the UEFA Euro 2004 Final was "Angelos Charisteas" with a header and not "Angelos Basinas" through a penalty kick. To tackle this problem, the fact-checking service first collects the facts from the ChatGPT response (in RDF format), and then through dedicated algorithms based on semantic web techniques, word embeddings and sentence similarity metrics, finds the most similar fact(s) in the RDF KGs indexed by LODsyndesis. The objective is both to confirm the correct ChatGPT facts and to find the correct answer for erroneous ChatGPT facts from existing RDF KGs. In the lower side of Figure 2, we can see that the fact-checking service managed to find the correct answer for the scorer of "UEFA Euro 2004 Final", that is, "Angelos Charisteas", and provided also the right provenance for this information (i.e. the DBpedia KG).

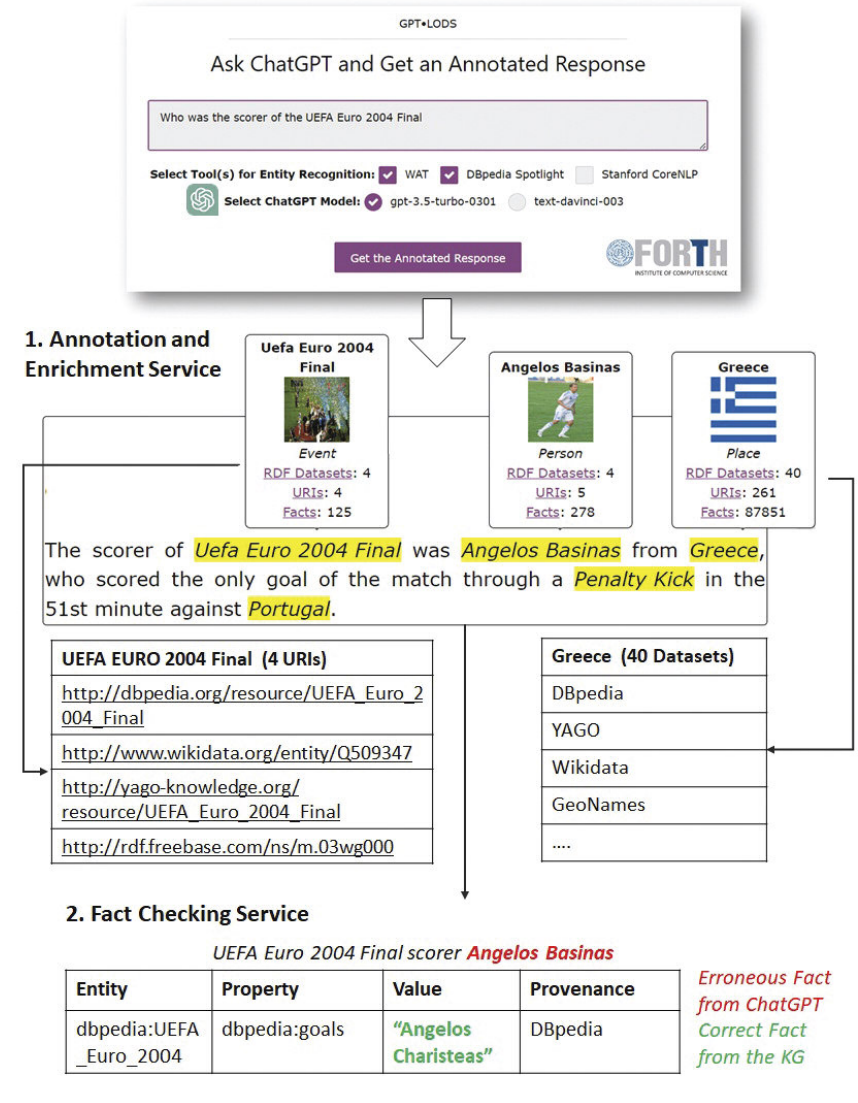


Figure 2: The services of GPT•LODS.

In the future we plan to extend GPT•LODS for supporting fact validation from more types of sources (including web pages), to evaluate the quality of the validation service and to provide a REST API, for enabling the exploitation of these services in various applications.

Links:

[L1] https://demos.isl.ics.forth.gr/GPToLODS

[L2] https://demos.isl.ics.forth.gr/lodsyndesis/

References:

[1] M. Mountantonakis, and Y. Tzitzikas, "Using multiple RDF knowledge graphs for enriching ChatGPT responses," In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2023), Demo Track. 2023.

[2] M. Mountantonakis, and Y. Tzitzikas, 2020. "Content-based union and complement metrics for dataset search over RDF knowledge graphs," Journal of Data and Information Quality (JDIQ), vol. 12, no. 2, pp.1–31, 2020.

Please contact:

Michalis Mountantonakis
FORTH-ICS and University of Crete
mountant@ics.forth.gr

Yannis Tzitzikas, FORTH-ICS and University of Crete
tzitzik@ics.forth.gr, +30 2810 391621,

An Intuitive Architecture for a Chatbot that Exploits High-Level Reasoning for Human–Robot Interaction

by Christoforos Prasatzakis, Theodore Patkos and Dimitris Plexousakis (ICS-FORTH)

In this article, we present a new, easy-to-understand and flexible chatbot architecture, which banks on ease of use and modularity. It relies on the Event Calculus, in order to perform high-level reasoning on world events and agents' knowledge, and can answer questions regarding the other agents' belief state, in addition to what is happening in the chatbot's world in general.

Human and computer interaction is a rather "hot" topic in modern computer science research. One of the many applications of human–computer communication technologies are chatbots, i.e., computer programs that engage in dialogue with humans, who ask questions in natural language and receive appropriate answers. The answers must not only be well formed and justified, but they must also be as close to the human's intuition as possible. And while we try to fulfil the above goal, more questions arise as to the chatbot's abilities. Such questions are: if the chatbot's world is inhabited by more than one agent, what actions have been performed by each and what the other agents have observed? What does each agent believe about the world? Also, how does past interactions and/or beliefs affect what the agents perceive and communicate now? And finally, is there an easy-to-understand and intuitive way for the chatbot to represent the knowledge it receives?

To address all of the above issues, we are developing a chatbot architecture – part of the SoCoLa project [L1] – that can handle the above tasks in an easy-to-understand-and-handle manner. Our architecture makes use of Event Calculus (EC) [1] – implemented in Answer Set Programming (ASP) [2] – in order to describe events and fluents (situations that "hold" for a certain time frame) in the chatbot's world. The architecture can use the state-of-the-art ASP reasoner Clingo [L2], in order to capture the dynamics of a given domain, infer causal relations, infer what has happened in past time points or what may happen in future time points in the chatbot's world, as well as answer user questions on hypothetical scenarios (of the "What would happen if…" type). The chatbot makes use of Meta's wit.ai [L3] natural language processing platform, in order to mark entities in user questions, as well as designate the question's intent and type.

The architecture consists of several distinct components as seen in Figure 1. Clingo and wit.ai are off-the-shelf tools we properly parameterise for our needs. The control/processing unit is written entirely in Python and serves as a "proxy" for all other components, while at the same time handling the burden of answering the user's question, and the graphical user interface (GUI) is built on Node.js, and sends the current world's state to the controller, as well as the user's question.

Our system is domain independent to the extent possible and supports a set of question types that we constantly expand. The current implementation supports: polar questions (e.g., "Did the human pick up the pen?"), which can be answered with either a "Yes" or "No" response; "where"-type questions, which ask the position of an agent or an object (e.g., "Where was the

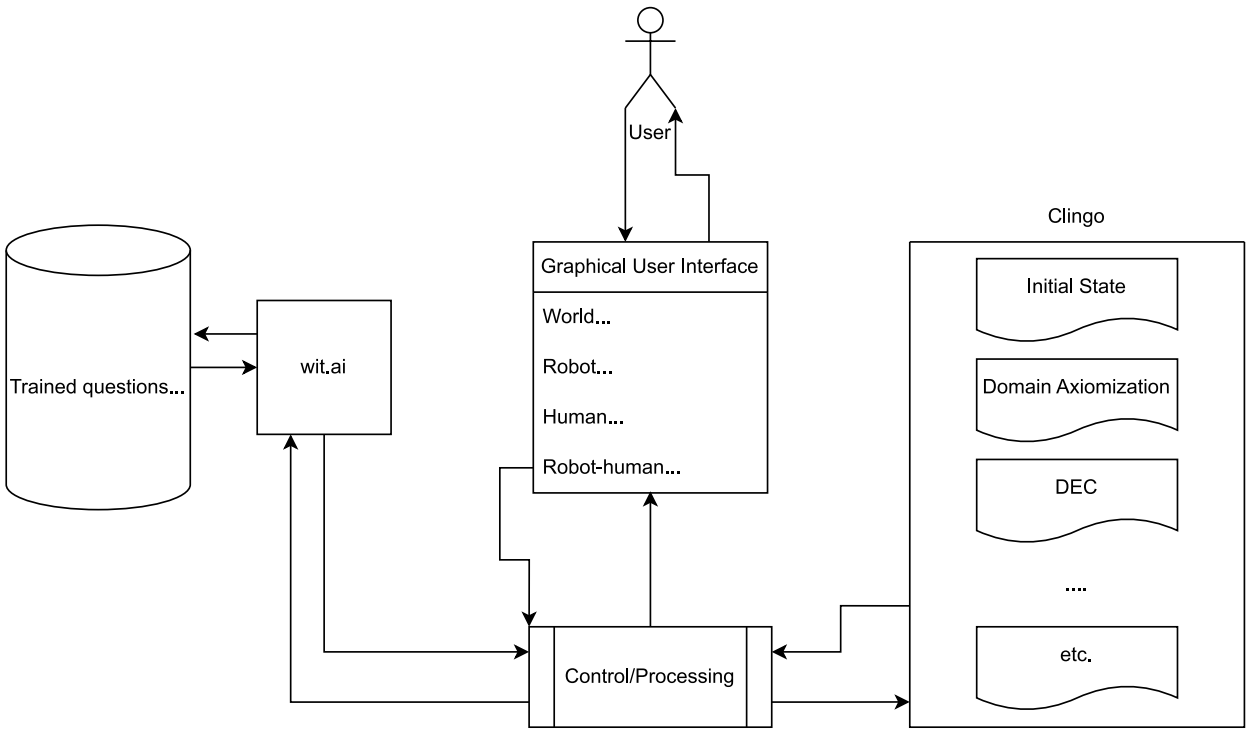


Figure 1: A block diagram showcasing the architecture and its workflow.

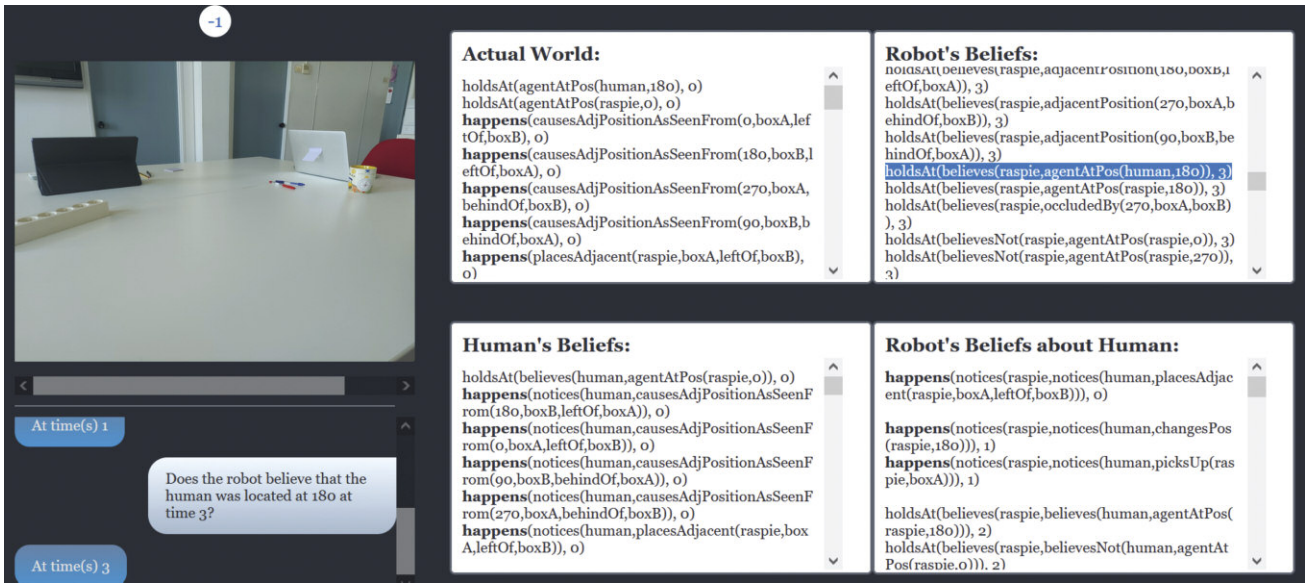


Figure 2: An epistemic question answered by our system.

robot located at time 3?"); "when"-type questions, which capture temporal aspects of the domain (e.g., "When was the robot located at angle 180"); and "what-if"-type retrospective questions, which ask what would have happened – in a hypothetical setting – if an event occurred at a given time point (e.g., "What if the human picks up the pen at time 0?").

A distinctive feature of our chatbot is that while it can answer typical questions based on what is taking place in the world, it also has the ability to answer questions on what individual agents observe or believe, both about the actual world or about other agents. This question type is called epistemic. Two examples of epistemic questions are "Did the human notice that the robot picked up the pen?" and "Does the human believe that the robot is located at angle 180?". Figure 2 shows an example of our system processing a sample epistemic question. In an environment, where more than one agents interact and modify the world, such questions can be of real value, since the agents may have incomplete or erroneous beliefs about that the current state of the world is or what has occurred. As such, understanding their beliefs can help explain better their behaviour.

The layout of the GUI is as follows: the top-left pane shows Event Calculus-encoded knowledge about the chatbot's actual world. The top-right pane shows the chatbot's knowledge about agent "robot". The bottom-left pane is the same for agent "human" and the bottom-right pane showcases what "robot" believes about "human", thus, allowing us to ask questions on an agent's knowledge about other agents, not just the environment.

The axiomatisation of a domain with an expressive formal theory, such as the Event Calculus, along with the ability to perform high-level reasoning on events that change the world or an agent's perspective about the world, enables the chatbot to respond with replies that are based on intuitive, well-justified conclusions. Causal, temporal, epistemic and retrospective aspects are all highly relevant for building intuitive human–robot interaction systems. "Why" questions, which we currently work on, will further enhance the explainability of our chatbot, by exploiting the structured, interpretable chain of conclusions made by the reasoner. The challenge there is to offer explanations that are both sound and minimal, in terms of relevance to what the user is interested in understanding.

Overall, our chatbot architecture offers a high degree of modularity. It also has great potential for expansion and improvement. Currently, it can support nested epistemic questions up to two levels. Future revisions may add further levels of nesting, as well as support for more question categories. Adding new question categories may open up the door for greater flexibility and applications, allowing the architecture to be used in even more areas than initially designed.

Links:

[L1] https://socola.ics.forth.gr/
[L2] https://potassco.org/clingo/
[L3] https://wit.ai/

References:

[1] R. Kowalski and M. Sergot, "A logic-based calculus of events," New Generation Computing, vol. 4, no. 1, pp. 67–95, 1986. doi:10.1007/BF03037383. ISSN 1882-7055. S2CID 7584513.
[2] V. Lifschitz, "What is answer set programming?," in Proc. 23rd National Conference on Artificial Intelligence, AAAI Press, 2008, 3: pp. 1594–1597.

Please contact:

Christoforos Prasatzakis, ICS-FORTH, Greece
xrprasas@ics.forth.gr

Theodore Patkos, ICS-FORTH, Greece
patkos@ics.forth.gr

A Pathway to Combat Climate Change with Human-Centred XAI

by Anahid Jalali, Alexander Schindler (AIT) and Anita Zolles (BFW)

Explainable Artificial Intelligence (XAI) is gaining importance in various fields, including forestry and tree-growth modelling. However, challenges such as evaluating model interpretability, lack of transparency in some XAI methods, inconsistent terminology, and bias towards specific data types hinder its integration. This article proposes combining long short-term memories (LSTMs) with example-based explanations to enhance tree-growth models' interpretability. By generating counterfactual examples and engaging domain experts, critical features impacting outcomes can be identified. Addressing privacy protection and selecting appropriate reference models are also crucial. Overcoming these challenges will lead to more interpretable models, supporting informed decision-making in forestry and climate change mitigation.

Explainable Artificial Intelligence (XAI) is a rapidly developing field that seeks to create justifications for machine learning models' decisions. Interpretable machine learning (ML) techniques such as decision trees, linear regressions and k-nearest neighbours have long been used to model complex phenomena and their associated data. However, data's growing size and complexity have led to the development of more sophisticated methods. Consequently, recent research has shifted towards using black-box models, typically deep learning models.

With the increasing popularity of these large and intricate models, the demand for explainability is also on the rise. This is particularly crucial in sensitive applications such as AI for forestry and tree-growth modelling, where decisions can have significant social and environmental consequences, and the models must be interpretable to support decision-making. Therefore, there is a growing need for explainable AI to provide transparency and accountability in decision-making processes.

However, there are still challenges with the existing approaches. Multiple studies have already reviewed the drawbacks of the current methods [1,2]. To understand the challenges associated with integrating XAI approaches, it is essential to have a grasp on the principles and techniques of XAI, as well as the organisation and content of data.

- Challenge 1. Evaluation of model interpretability: Multiple studies have approached this challenge with user-centric methods, proposing quantitative and interpretability metrics to evaluate the effectiveness of XAI methods. However, a unified metric for assessing the quality of the explanations is still an open question.
- Challenge 2. Lack of transparency of some XAI methods: Some use complex algorithms, such as neural networks, to generate explanations. While these methods can provide accurate explanations, it may not be easy to understand how they arrive at their conclusions.
- Challenge 3. Inconsistent terminology: Researchers and practitioners may use different terms and definitions for similar concepts, leading to confusion and misunderstanding.
- Challenge 4. Existing bias in developing XAI methods: There is a bias towards computer vision, tabular data, and natural language processing, as well as a need for more research for appropriate XAI, approaches for other complex data such as multivariate time series, and geospatial data.
- Challenge 5. Integration of domain knowledge into XAI methods: Domain knowledge can provide crucial insights into the underlying data-generating processes and the context in which the data is collected, which can help enhance the model's interpretability and accuracy.
- Challenge 6. Explanation's coverage concerning the underlying black-box model: This challenge addresses the provided explanations, which should accurately reflect the underlying black-box model. In other words, the explanations should provide sufficient coverage of the model's decision-making process to enable users to understand the rationale behind the model's predictions.
- Challenge 7. Lack of consideration for social and ethical aspects within XAI: The social and ethical aspects of XAI are often overlooked, and it is crucial to adapt privacy-preserving methods to ensure that XAI considers the necessary privacy-protection measures for mobility data.

The use of AI for forestry and tree-growth modelling has emerged as a promising tool to mitigate the impact of climate change on forests. AI algorithms can help predict the future

Figure 1: Protecting and managing forests with AI. Image source: Karlsruhe Institute of Technology.

Figure 2: An automatic dendrometer, a high-resolution measuring device that records the tree circumference change every hour.

growth of trees and their response to climate change factors using time series data that capture temperature, rainfall and carbon dioxide levels. With this information, we can identify the best tree species for planting, optimise forest-management practices, and reduce carbon emissions through sustainable forest management [3]. Moreover, explainable AI can provide transparency and accountability in decision-making, making adopting and implementing climate-friendly policies easier.

While the challenges in applying XAI to forestry and tree-growth modeling are similar to those in other fields, there are also unique hurdles that need to be addressed, such as the requirement for more data on certain aspects of tree growth, including root development and tree competition factors. In order to tackle these obstacles, our research plan for AI in tree growth involves creating appropriate models and XAI techniques.

Long short-term memories (LSTMs), a variant of recurrent neural networks, show promise as a deep learning model for multivariate time series modelling. Recent studies indicate that LSTMs effectively capture temporal dependencies and relational information in time series data. As a result, LSTMs are increasingly used to accurately model and predict complex phenomena such as tree growth.

Our research hypothesis is that using example-based explanations is more effective in explaining model decisions to end-users than gradients and heatmaps, which are more useful for ML developers. We believe that counterfactual examples, generated by manipulating attributes such as weather and soil parameters, can help understand the impact of temporal changes in the past. By observing how the model's predictions differ under alternative conditions, we can identify critical features contributing to specific outcomes and enable stakeholders to make informed decisions based on the model's explanations.

To enhance the explainability of predictive models for tree growth, we will combine LSTMs and example-based explanations. We will engage domain experts to develop and evaluate the required temporal visualisations and integrate temporal semantics. Moreover, using such explanations, we can identify errors and biases in the data, and with the feedback of domain experts, we can increase the data quality and, therefore, the model performance.

Additionally, privacy-preserving methods will be adapted to ensure that XAI accounts for the privacy protection necessary for data owners and model robustness against adversarial attacks. Selecting the appropriate reference data and models is a significant challenge in XAI. Our approach includes evaluating the effectiveness of various reference models and selecting the best ones for each specific case. By addressing these challenges, we hope to develop more interpretable models that can be trusted by both experts and end-users, leading to better-informed decisions in forestry and tree-growth modeling.

References:

[1] A. Theissler, et al., "Explainable AI for time series classification: a review, taxonomy and research directions," IEEE Access, vol. 10, pp. 100700-100724, 2022, doi: 10.1109/ACCESS.2022.3207765.

[2] U. Schlegel, et al., "Towards a rigorous evaluation of XAI methods on time series," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 4197–4201, doi: 10.1109/ICCVW.2019.00516.

[3] J. Olivar, et al., "The impact of climate and adaptative forest management on the intra-annual growth of Pinus halepensis based on long-term dendrometer recordings," Forests, 2022; vol.13(6):935. https://doi.org/10.3390/f13060935

Please contact:
Anahid Jalali, AIT Austrian Institute of Technology, Austria
anahid.jalali@ait.ac.at

The Eye of the Beholder Project: Transparent and Actionable AI Pipelines for Information Quality Prediction

by Davide Ceolin (CWI) and Ji Qi (Netherlands eScience Center)

The spread of disinformation affects society as a whole and the recent developments in AI are likely to aggravate the problem. This calls for automated solutions that assist humans in the task of automated information quality assessment, a task that can be perceived as subjective or biased, and that thus requires a high level of transparency and customisation. At CWI, together with the Netherlands eScience Center, we investigate how to design AI pipelines that are fully transparent and tunable by an end user. These pipelines will be applied to automated information quality assessment, using reasoning, natural language processing (NLP), and crowdsourcing components, and are available in the form of open source workflows.

The assessment of the information quality of online content is a challenging yet necessary step to help users benefit from the vast amount of information available online. The challenge lies in the complexity of the assessment process, the expertise and resources required to perform it, and the fact that the resulting assessment needs to be understood by laypeople to be trusted and, therefore, used. This means that, on the one hand, we need to scale up the process, which is onerous, while the number of items to analyse is vast. Scalability is essential to guarantee the feasibility of the effort. On the other hand, we need to aim at explainable and transparent approaches to address this problem. For the user to trust the result of the computation, they need to understand the computation process and the quality assessment itself should be explainable.

The Eye of the Beholder project [L1] aims to enhance the transparency and the explainability of information quality

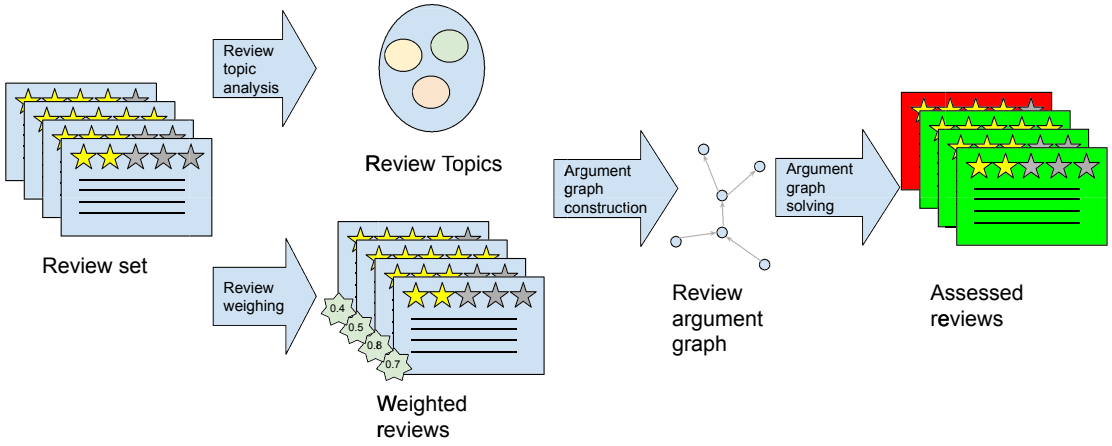


Figure 1: Example of an argument-based product review assessment pipeline.

assessments. The Eye of the Beholder started in 2022, is led by the Human-Centered Data Analytics Group [L2] at CWI and is funded by the Netherlands eScience Center. We focus on one specific class of information items, namely product reviews. Reviews are meant to provide second-hand assessments about the quality of products; however, their quality is diverse, and since they represent personal opinions, they are difficult to verify.

Several product-review collection websites, however, address the problem of reviewing the quality of reviews by collecting judgements from the readers, for instance in terms of upvotes (e.g., 'likes' or 'thumbs ups' collected to rank reviews). This method helps spot the most useful, informative, and possibly high-quality reviews. Previous work of ours [1] showed that employing an argument-based approach is a promising way to automate the identification of the most useful product reviews. This is useful because it helps identify high-quality reviews before users spend time on them. However, mining arguments in text and reasoning on their strength is a difficult task. Several AI components can be employed in order to identify arguments based on the sentence structure, text topic and text similarity, and these arguments can be evaluated against conflicting ones by employing different reasoning and different weighing schemes. When conflicting arguments are identified, it is possible to indeed weigh them to determine their quality, but there is no universal method to do so. Some of the models that allow reasoning on these arguments, for instance, allow modelling of support among arguments explicitly, while others focus on modelling attacks only. In principle, both types of model could serve our purposes but, in practice, we need to evaluate which of them better fits the use case at hand.

For this reason, within Eye of the Beholder, we developed a series of open source workflows that implement such computational argumentation-based information quality assessment pipelines. These workflows have been implemented in the Orange platform [L3]. These open source platforms allow design AI and machine learning workflows, and by leveraging them, we enhance the transparency of our pipeline for the purpose of actionability. This means that we do not only make the pipeline transparent and understandable by users, but we also allow users to tune and tweak the pipeline itself. For instance, users may want to increase or decrease the granularity of their topic-detection component (so to allow it to identify coarser- or finer-grained topics), or, it may want to use a different argument-reasoning engine than the one we propose. By using our workflows, the user will be able to make these changes and evaluate their performance implications, or they can also propose new theoretical frameworks to this purpose and evaluate them. For instance, in our project, we developed a theoretical framework for reasoning specifically on the arguments of product reviews [2] and we will evaluate it in the future through our workflows.

In the future, we will explore the development of these automated pipelines further. While implementing them, we are focusing on determining methods to anticipate the influence of parameter tuning on the pipeline performance. This will help us guide non-technical users in the management of the computation.

Lastly, we collaborate with the AI, Media and Democracy lab [L4] – an interdisciplinary lab that brings together AI researchers, communication scientists, information law scholars and media organisations that will provide an ideal setting to evaluate applications of our approach in the media and democracy fields.

Links:

[L1] https://kwz.me/hjN
[L2] https://kwz.me/hxy
[L3] https://kwz.me/hxH
[L4] https://www.aim4dem.nl/

References:

[1] D. Ceolin et al., "Transparent assessment of information quality of online reviews using formal argumentation theory," Information Systems, vol. 110, 2022, Art. no. 102107.
[2] A. K. Zafarghandi and D. Ceolin, "Fostering explainable online review assessment through computational argumentation," in Proc. 1st International Workshop on Argumentation for eXplainable AI, ArgXAI@COMMA, 2022.

Please contact:

Davide Ceolin, CWI, The Netherlands
davide.ceolin@cwi.nl

Explainable AI for Astronomical Images Classification

by Mahmoud Jaziri (Luxembourg Institute of Science and Technology) and Olivier Parisot (Luxembourg Institute of Science and Technology)

AI is an indispensable part of the astronomer's toolbox, particularly for detecting new deep space objects like gas clouds from the immense image databases filled every day by ground and space telescopes. We applied explainable AI (XAI) techniques for computer vision to ensure that deep sky objects classification models are working as intended and are free of bias.

Recently, deep neural networks became state of the art in many fields, outperforming domain experts in some cases. With EU regulation (GDPR and the future AI Act), explainable AI (XAI) has become a hot-topic issue. It is also important for the scientific community. Whether it is classifying new quasars or detecting new gas clouds, astronomers need AI to automatically process millions of deep sky images. But how can we be sure of an AI model's accuracy? And how can we prove that there is no bias in the data, or the implementation?

We have tested classification methods on astronomical images during the MILAN project, funded by the Luxembourg National Research Fund (FNR), grant reference 15872557. To this end, we have trained models to classify deep sky objects (galaxies, nebulae, stars clusters, etc.). Based on VGG16 and ResNet50 architectures, we have applied these models on images captured with smart telescopes [1].

Two XAI approaches were investigated: a global one attempting to explain the entire AI model's decision-making process, and a local one that explains a single prediction. Each approach satisfies a certain need in the XAI debate. Respectively offering insight into which features are overall most important (e.g. for predicting heart disease); or getting into the details of a single prediction, offering insight into which features are most important for the investigated output. This is the most used approach in computer vision [2] and the one that we will be discussing.

Some techniques attempt to trace a neural network's output back through its layers and gauge its "sensitivity" to certain image features. These methods are appropriately called gradient-based methods, e.g. guided backpropagation (GBP), integrated gradients (IG), concept activation maps (CAM), layer-wise relevance propagation (LRP) and DeepLIFT (Deep Learning Important FeaTures). Some try to selectively perturb a neural network, by adding noise to the input, modifying or removing subsets of neurons or features, then recording the network's output and inferring their importance. This leads to the construction of an "attribution map" on which the most important features have higher importance scores. These perturbation-based methods are often combined with gradient-based techniques to produce two of the most popular techniques used:

- SHAP (SHapley Additive exPlanation) assigns importance to features based on their average contribution to every possible subset of features.
- LIME (local interpretable model agnostic explanations) creates an interpretable simpler model, called surrogate, such as a logistic regression or a decision tree to serve as a local approximation to the original model, reducing the difficulty of interpretation enough for human understanding.

Individually these techniques are inherently flawed, and the extent of their validity is limited [3]. Some methods are blind to a feature with maxed-out contribution, i.e. locally, the output is not "sensitive" to the feature, therefore it is considered unimportant despite its criticality for the correct prediction. Other methods produce different attribution maps for functionally equivalent neural networks. Most are vulnerable to manipulation, mapping on the same image – the same explanation for different outputs, or different explanations for the same output prediction. Furthermore, they are heavily dependent on hyper-parameters (e.g. the number of samples) and the training data biases.

Overall, the discussed techniques provide explanations that are too brittle and could lead to a false conclusion about the model's performance. Solving this problem has recently taken the XAI community into an exciting new direction researching the robustness of AI explanations. It posits five axioms and their mathematical formulations to evaluate the quality of an explanation technique by measuring its ability to deal with the discussed limitations. Respectively to the mentioned flaws, these axioms are: saturation, implementation invariance, fidelity, input invariance, and sensitivity [3]. Though there is no consensus on which XAI technique best satisfies these axioms, the trend is to mix and match different techniques making them inherit each other's robust qualities. This has led to new techniques such as LIFT-CAM, Guided CAM, Smooth IG, etc.

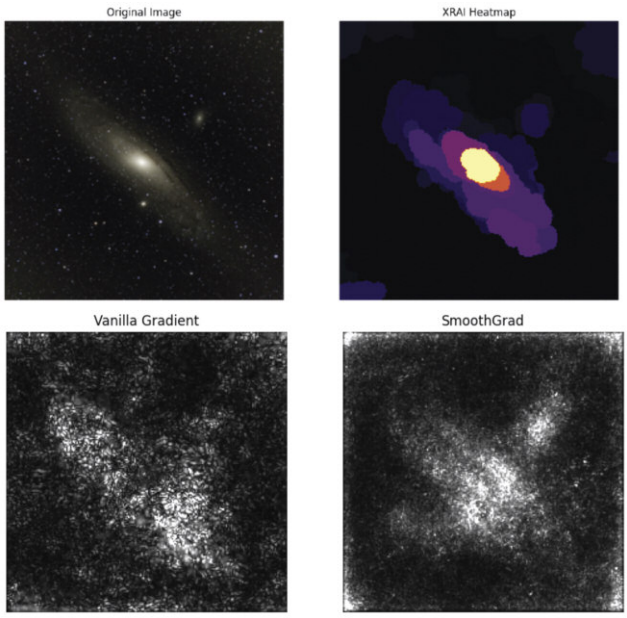


Figure 1: the first image represents the Andromeda Galaxy (M31) captured with a smart telescope (top left). The other images are the output of different approaches to explain the detection of deep sky objects with a TensorFlow VGG16 classifier.

The robustness of an explanation is measured by evaluating the similarities between two attribution maps: a reference attribution map from an unperturbed image, and that of a perturbed image.

Evaluating the two images relative to each other gives us a great overview of the performance of our explanation method on the defined five axioms. To do so, several visual metrics like SSIM (structural similarity index measure) help in this evaluation, resulting in explanations closely coherent with the human visual intuition [2]. In fact, the robustness of an explanation and the model itself are closely related, e.g. the non-smooth decision boundaries of a neural network make its attribution map vulnerable to small natural perturbations in input. In an ideal world, the attribution maps should be sensitive enough to detect noise but immune to their effects.

When applying these XAI techniques on our deep sky objects classification models, we obtained attribution maps which highlight the image regions considered most important by the classification model (see Figure 1). Vanilla Gradient shows the pixels the model is sensitive to, and SmoothGrad denoises the attribution map to moderate use (two clusters for the two galaxies). XRAI's segmentation method produces better results, accurately detecting the centre and shape of the bigger galaxy, but less so the second smaller galaxy. However, across the methods used, there is a small unwanted bias in the model to use low-light corners. In future works, we will continue to refine our classification model architectures to find a satisfying trade-off between accuracy and interpretability.

Link:

[L1] https://www.list.lu/en/news/high-quality-noise-free-astronomical-images/

References:

[1] O. Parisot, et al., "MILAN Sky Survey, a dataset of raw deep sky images captured during one year with a Stellina automated telescope," Data in Brief, vol. 48, 2023, Art. no. 109133.

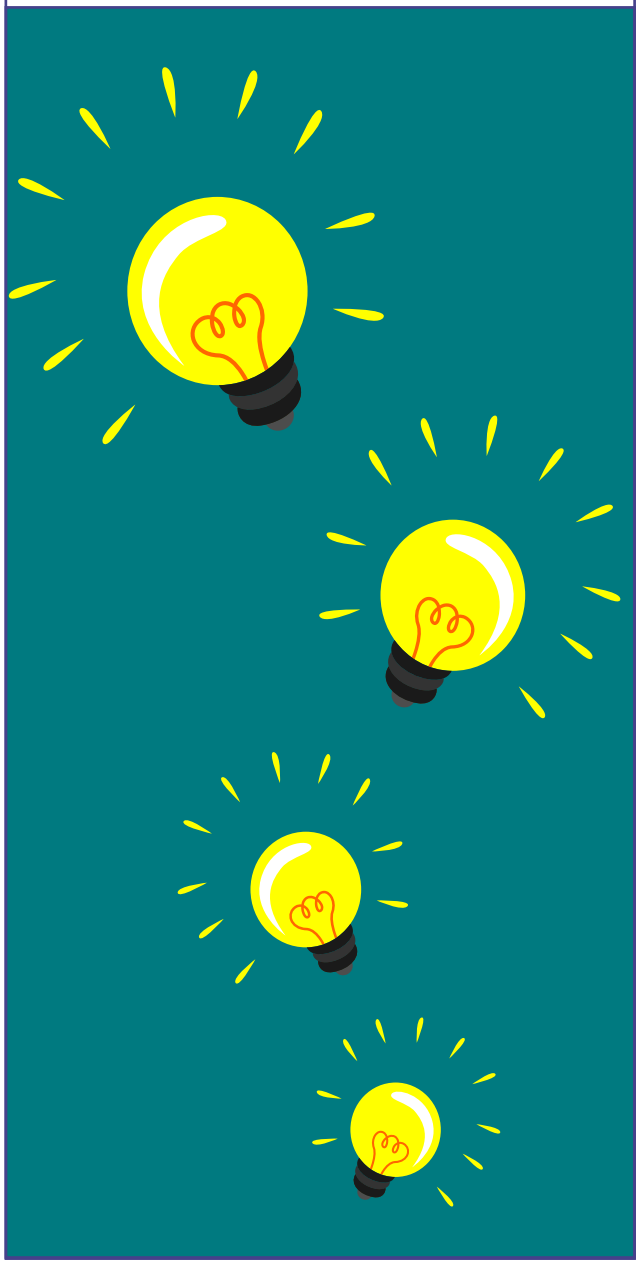
[2] M. V. S da Silva, et al., "Explainable artificial intelligence on medical images: a survey," arXiv preprint arXiv:2305.07511, 2023.

[3] I. E. Nielsen, et al., "Robust explainability: a tutorial on gradient-based attribution methods for deep neural networks," in IEEE Signal Processing Magazine, vol. 39, no. 4, pp. 73–84, Jul. 2022, doi: 10.1109/MSP.2022.3142719.

Please contact:

Mahmoud Jaziri, Luxembourg Institute of Science and Technology, Luxembourg
jazirimahmoud360@yahoo.com

Olivier Parisot, Luxembourg Institute of Science and Technology, Luxembourg
olivier.parisot@list.lu



European Research and Innovation

Fibre-Optic Sensing for Road-Traffic Monitoring in Remote Areas

by Martin Litzenberger, Carmina Coronel and Kilian Wohlleben (AIT Austrian Institute of Technology GmbH)

Accurate real-time traffic sensing is of key importance for optimising and managing road traffic. Often the high density of traffic sensors, needed to achieve an accurate real-time monitoring of important roads, is difficult to implement due to technical constraints or because of high installation cost. Fibre-optic sensing (FOS) is a new and cost-effective alternative technology that allows a seamless, real-time monitoring of the road traffic over large distances of up to 50 km, even in remote areas such as on critical costal or mountain roads, using existing telecom fibre-optic cable infrastructure.

Even with an upcoming transition to electric mobility and a modal shift from individual to public transport in the future, roads will stay the backbone of transportation for years to come. Therefore, permanent traffic monitoring is crucial to ensure optimal traffic flow. The data provided by real-time road-traffic monitoring provides information regarding traffic jams or accidents. With such information, traffic-management centres are enabled and supported to react quickly to incidents and intelligent transportation system (ITS) measures, for example, the closure of a lane or temporary usage of the hard shoulder, can be imposed. Often the large number of traffic sensors, needed to achieve an accurate real-time monitoring of important roads, is difficult to implement due to technical constraints or because of high installation cost. Furthermore, in remote areas where traffic monitoring can be also of importance (e.g. on difficult and narrow road sections in costal or mountain roads), a dense coverage and easy connectivity of road-traffic sensors is problematic. One method for traffic monitoring is through crowdsourcing of smartphone connection data [1] or from fleets of vehicles equipped with GPS systems ("floating car"). Google Maps is the most prominant example of the crowdsourcing approach. However, it is unable to deliver true real-time information, it relies on traffic models, and it needs the "cooperation" of the data providers, that is, a high enough number of mobile phone users in the area of interest. In addtion to that, mobile phone and wireless data connectivity might also be reduced in remote areas.

Fibre-optic sensing (FOS), also often termed "distributed acoustic sensing" (DAS) [2], is a technology that allows a seamless, real-time monitoring of vehicle trajectories on a road over large distances of up to 50 km without additional roadside installations. It uses one unused optical fibre of an optical cable already installed in the ground for data and communication networks (telephone, Internet), as a distributed sensor. With increasing "fibre-to-the-home" initiatives througout Europe, even in remote areas, fibre-cable infrastructure will become more availble, and will be typically installed next to roads.

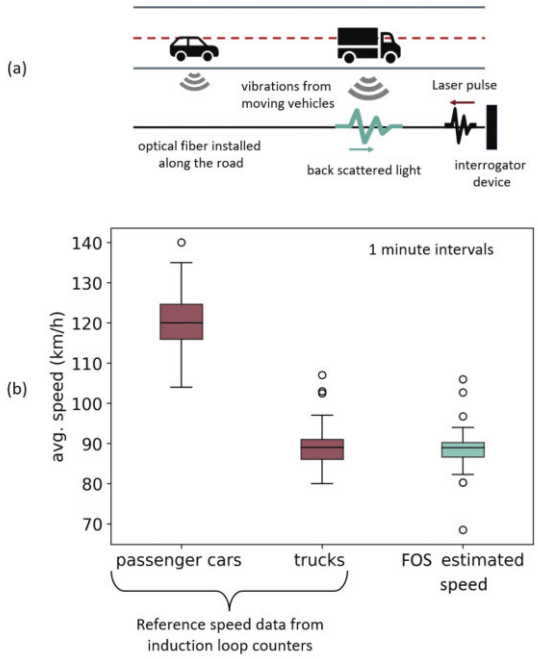


Figure 1: (a) Principle of FOS road-traffic monitoring. The fibre optic cable picks up vehicle vibrations that are probed via the back-scattered light of the interrogating laser pulse. (b) Comparison of induction-loop counter-reference-speed measurements (cars and trucks separated) and the accumulated average speed estimated from the FOS measurement (no vehicle classification).

In FOS systems, an interrogator device connected to one end of the fibre transmits a series of laser light pulses into the fibre-optic cable, then portions of the light pulses are back-scattered from inside the fibre and are measured by the interrogator. Any vibrations generated by road traffic nearby the cable stretch and compress the optical fibre on a microscopic scale, affecting the optical-path length. An interferometer in the interrogator unit measures time of flight and the phase shift of the back-scattered light and thereby determines the position and trajectories of road vehicles (Figure 1a).

We have developend and applied a FOS traffic-flow-sensing algorithm [3] over a distance of 19 km with the fibre-optic cable being installed in parallel to a highway in a mountain region in Austria, and compared the results to the speed measurements from a reference induction-loop detector installed in the road surface of the same road.

The FOS signal is a combination of signals originating from passenger vehicles and trucks for both lanes of the highway. The current detection algorithm does not distinguish between vehicle classes nor lanes. In line with this, we performed the comparison by comparing our results to both accumulated truck and passenger vehicle results from the reference detector. The box plots (Figure 1b) show the average speeds of the passenger cars (ground truth) and trucks (ground truth) in 60 one-minute intervals, compared to the average speeds estimated by the FOS algorithm, in the same time intervals, at the same position along the road.

The reference detector shows that the speed of the measured trucks (n = 94) is in the range of 80–100 km/h, which is slower than the passenger vehicles (100–135 km/h). The estimated average speeds extracted by the FOS algorithm show that the FOS results are in good agreement with the average speed of trucks, suggesting that the cars are not picked up by the FOS measurement.

Passenger cars and vehicles on the opposite carriageway, away from the fibre-optic cable, were not detected reliably. In detail, we found that: (i) passenger cars were not detected if the signal intensities were not strong enough, because the roads' traffic lanes were too far away from the fibre cable (up to 20 m distance in the described setup), (ii) the algorithm failed when vehicles were driving exactly in parallel or very close to each other, thus preventing our algorithm from correctly identifying the two distinct trajectories, or (ii) weaker signals originating from passenger vehicles have been masked by stronger signals originating from heavier and larger vehicles.

In conclusion, we found that the FOS system is capable of monitoring traffic situations based on average speed measurement but cannot reliably resolve smaller passenger cars. Also, the distance of the fibre-optic cable from the traffic lanes is crucial, and thus not all lanes of wide roads can be monitored. Therefore, we see the potential application of FOS traffic monitoring on, for example, smaller roads in remote areas, such as on costal or mountain roads, where a dense coverage and easy connectivity of conventional road-traffic sensors is problematic.

References:

[1] M. Lewandowski, et al., "Road traffic monitoring system based on mobile devices and bluetooth low energy beacons," Wireless Communications and Mobile Computing, vol. 2018, pp. 1–12, Jul. 2018. [Online]. Available: https://doi.org/10.1155/2018/3251598

[2] D. Hill, "Distributed acoustic sensing (DAS): theory and applications," in Frontiers in Optics, 2015.

[3] C. Wiesmeyr, et al. "Distributed acoustic sensing for vehicle speed and traffic flow estimation", in 2021 IEEE international intelligent transportation systems conference (ITSC), 2596–2601, 2021. Available at: https://doi.org/10.1109/ITSC48978.2021.9564517

Please contact:

Martin Litzenberger, AIT Austrian Institute of Technology GmbH, Austria
martin.litzenberger@ait.ac.at

Security-by-Design IoT Development and Certification with IoTAC

by Sascha Hackel, Martin Schneider and Ramon Barakat (Fraunhofer FOKUS)

The Horizon 2020 European research project IoTAC [L1] enhances IoT security through a multi-layered approach using best practices, standards and research findings. Its technology includes privacy-friendly access-control mechanisms, ML-based attacked detection, and security-by-design methodology. Additionally, IoTAC provides developers and service providers with SDKs and APIs to seamlessly integrate the framework.

The IoTAC project [L1] complements innovative security features to enhance the Internet of things (IoT) security at both the architecture and application levels based on the ISO/IEC 30141 IoT reference architecture [L2]. It integrates security, safety, resilience, trust and privacy as cross-domain functions and has two development tracks: one extends the IoT architecture with functional features, and the other designs the IoTAC Software Security by Design and Security Assurance Model platforms. Both tracks meet in the assessment and validation procedure of IoT platforms and software applications.

Functional Security Modules

The IoTAC project develops several runtime modules to enhance the security of IoT systems. One of these modules is the front-end access management, which includes a privacy-friendly user authentication and authorisation using secure elements such as chip cards and secure tokens. This approach enables a decentralised operation, with user credentials stored in a secure element for improved privacy and security. The architecture also supports real-time issuance and management of credentials, making it easier to modify and revoke them as needed.

Another module is the IoT-enabled honeypots, used to attract potential attackers and monitor their behaviour to understand attack patterns and adopt appropriate security measures. The honeypot consists of two layers, one visible to attackers and another dedicated to analysing their behaviour. The system uses lightweight and advanced anomaly-detection techniques to detect behavioural changes in IoT devices to identify potential intrusions.

The security gateway module is enhanced with attack detection and prevention (ADP) and checkpointing features. The ADP mechanisms are developed based on deep learning, emphasising the adoption of advanced algorithms such as deep dense random neural networks to achieve better predictive performance with lower computational resources. Checkpointing is used to introduce new security countermeasures and restart the protected system from a safe operating point established at the most recent checkpoint.

The runtime monitoring system (RMS) collects security-related data from monitored IoT components or applications in real-time and stores it for further processing. The collected data is used to detect patterns of abnormal behaviour through analytics algorithms. The RMS features lightweight monitoring probes responsible for data collection and publishing to the monitoring platform. The probe management is facilitated by an internal probe registry that maintains probe information and status and enables probe creation, reconfiguration and discovery.

Finally, the project also includes an AI-based attack detection module that uses the auto-associative random neural network (AARNN) to provide highly accurate attack detection of major botnet attacks. Botnet attacks pose a significant threat to IoT systems, inducing compromised nodes and taking down specific nodes and links between devices and servers that process data. This technique's training protocol relies on normal traffic patterns, without requiring data regarding all possible attack patterns that the network may encounter, making it more efficient than other attack-detection approaches.

The extended ISO/IEC 30141 domain-based reference model is presented in Figure 1, where newly introduced IoTAC functional modules are introduced.

Testing Phase Methodology

As cybersecurity threats continue to evolve, software-development teams need to prioritise security testing early in the software-development life cycle. Many security breaches in software systems are due to security vulnerabilities that originate in the implementation or third-party libraries used.

The European IoTAC project is incorporating test and security strategies into the DevSecOps life cycle, an extension of the DevOps life cycle. Functional tests are executed in the DevOps pipeline based on the functional description of the application to evaluate functionality and detect faults in implementation early.

To identify a suitable set of test cases, methods and techniques, random testing, equivalence partitioning, boundary value analysis and model-based testing can be used. While functional tests ensure that software behaves as it should, security testing is needed to examine systems or applications for existing vulnerabilities.

Security testing requires the definition of dedicated security requirements, derived from multiple sources, including regulatory compliance or organisational security policies, risk analysis, and security guidelines and standards, like the OWASP Application Security Verification Standard (ASVS) [L3].

Two approaches can be used to identify security vulnerabilities: static application security testing (SAST), and dynamic application security testing (DAST). SAST is a white-box approach that analyses source code for known vulnerability patterns, while DAST is a black-box approach that tests an application in its running state by executing simulated attacks.

The IoTAC project aims to integrate both SAST and DAST, in addition to functional testing, into the DevSecOps pipeline to detect vulnerabilities early and at lower costs, leading to a more secure system before deployment. The specification of the described approach is also published at European level by the European Telecommunications Standards Institute (ETSI)

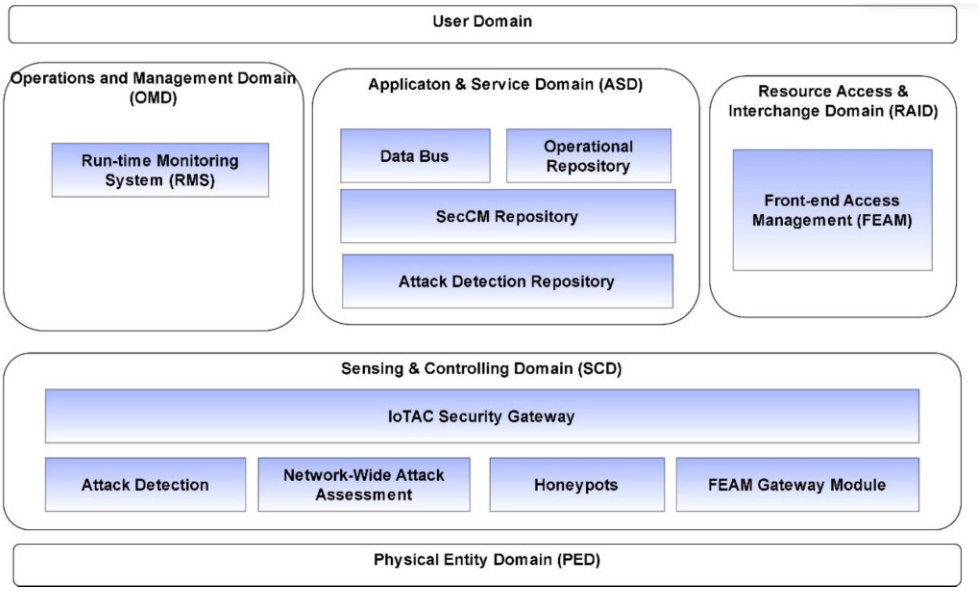


Figure 1: Extended ISO/IEC 30141 reference model (high-level view).

within the Technical Committee (TC) Methods for Testing and Specification (MTS) [L4].

By prioritising security testing early in the software-development life cycle and integrating automated security testing into the DevSecOps toolchain, the IoTAC project aims to provide and establish processes for the successful deployment of security-by-design approaches, increasing the security of IoT applications.

Security Validation Process

The IoTAC project seeks wide adoption of its architecture and platform, with a security-validation process that includes conformance and security testing. Figure 2 illustrates the proposed security validation process, which follows the EN 17640 standard for cybersecurity evaluation methodology of ICT products [L5]. This standard provides a minimum set of evaluation activities and evidence required for certification, reducing the burden on manufacturers.

The security-validation process involves three roles: the manufacturer who initiates the process, the evaluation facility that performs the assessment, and the validation authority that makes the certification decision based on the evaluation results. The IoTAC project aims to contribute to the European Cybersecurity Skills Framework, specifically to the development of an IoT cybersecurity certification scheme as part of ENISA's future roadmap.

The IoTAC project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 952684.

Links:

[L1] https://www.iso.org/standard/65695.html
[L2] https://iotac.eu/
[L3] https://kwz.me/hxS
[L4] https://kwz.me/hx4
[L5] https://kwz.me/hx0

Please contact:
Sascha Hackel, Fraunhofer FOKUS, Germany
sascha.hackel@fokus.fraunhofer.de

Martin Schneider, Fraunhofer FOKUS, Germany
martin.schneider@fokus.fraunhofer.de

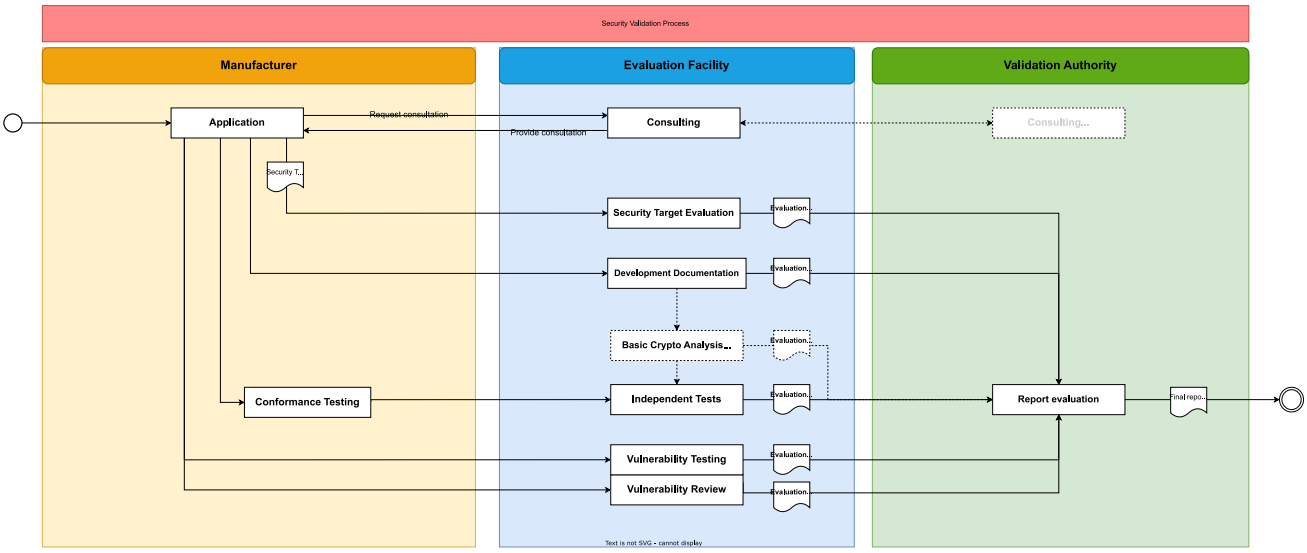


Figure 2: Proposed security-validation process.

SD4MSD: Using a Single Device for Multiple Security Domains

by Florian Skopik, Arndt Bonitz, Daniel Slamanig (Austrian Institute of Technology), Markus Kirschner (MUSE Electronics GmbH) and Wolfgang Hacker (Austrian Ministry of Defence)

Military field operations place high demands on information and communication technology (ICT) devices, both in terms of reliability and security. These requirements include robustness against environmental influences such as vibrations, water and humidity, as well as protection against physical attacks and cyberattacks [1]. Attempts to compromise a device must be detected immediately and, if necessary, trigger automated countermeasures such as alarms, partial deactivation or emergency wiping of all data.

Currently, there are robust end-devices from well-known non-European manufacturers in tablet form factor available on the international market. However, none of these devices support reliable monitoring of device integrity, nor do they combine all relevant international protection classes required for military use. In the interest of reusability, further measures must be taken to ensure that devices that were already used in one mission can be reliably and securely deployed in other security environments as new devices.

In contrast to other manufacturers, MUSE Electronics GmbH and its partners do not repurpose existing products, but develop an independent cyber-physical architecture for a robust computer tablet. A prototype has been created based on industrial design and reinforced plastics, which is also resistant to electromagnetic attacks. In addition, further hardware and software measures are implemented to harden the device, such as hardware security gateways that control the data flows between the components, and the use of cryptography to guarantee strong authentication between components and the integrity and authenticity of software running on the platform.

The SD4MSD project aims to combine an integral overall concept at physical, hardware and software level for highly robust end-user devices with individual configurability for specific (military) purposes. A comprehensive concept for hardening a mobile ICT device is developed in order to resist a sufficient range of physical, electromagnetic and cybersecurity-oriented attack vectors. Ensuring authenticity, integrity and confidentiality throughout the life cycle of the ICT device is crucial. A modular system architecture offers multiple purposes of use with simultaneous platform flexibility. Maintenance and service processes as well as best practices and standards are appropriately considered.

One Device for Multiple Security Domains

If the same devices are to be deployed at often short intervals one after the other in different mission-specific security domains, for example in command posts and mobile data centres, the security requirements increase significantly. Ensuring that highly sensitive and mission-critical data is only accessible in the currently relevant security domain is essential. International security and abuse incidents due to inadequate solutions based only on virtualisation and pure container solutions [2] underline the importance of a robust and reliable "single-device solution" that can be used in multiple application domains without security concerns. For security reasons, it is currently often necessary to purchase multiple physical devices to be able to change the device for each use, depending on the security domain for which it was ultimately certified. This is not only economically inefficient (such special devices, including expensive accessories, often incur high additional costs), but also poses a logistic challenge and causes additional work in the administration of devices. Hence, SD4MSD delivers a concept for a device that can be used in multiple security domains, isolating mission-specific data from each other without the risk of data spillover.

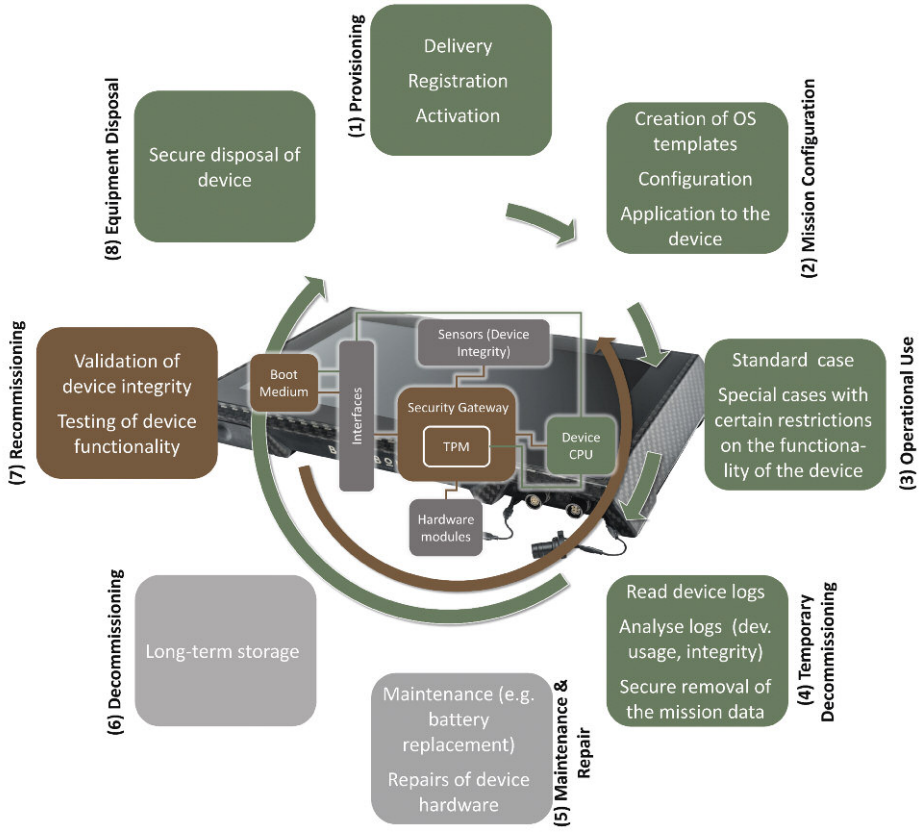


Figure 1: The SD4MSD life cycle model.

In the centre of Figure 1, the basic architectural elements of such a secure device are depicted. All mission data and the operating system (OS) is contained on an encrypted boot medium. A security gateway (SG) is built into the device, which takes over various security-critical tasks and plays a central role in the

concept. It is responsible for the control (activation and deactivation) of external and internal interfaces based on configuration files, communication with external media (read/write), and the implementation of a secure boot process. Furthermore, it monitors the physical integrity via sensors built into the device, logs derived events and may perform actions such as blocking or wiping in the event of an integrity violation. The SG also has access to Trusted Platform Modules. A TPM is a secure element that can generate and store cryptographic key information and perform cryptographic operations. Moreover, a TPM is central for guaranteeing platform integrity, that is, it can be used to store and verify the status configuration of software.

A Life Cycle Model for Reusable Handheld Devices

The project accounts for all phases in the life cycle of the secure tablet, as depicted in the outer elements in Figure 1, inspired by common IoT life cycle models [3]. This cycle starts with the provisioning of a device, including the initial delivery, the registration and activation of devices, and continues with mission configuration (i.e. the creation of mission-specific OS templates and deployment of applications), the actual operational use in the field and the following temporary de-commissioning once a mission has been completed. This last step takes care of reading out device logs, analysing them to detect any hardware or software issues, including compromise, and addresses the secure and permanent removal of mission-specific data. A device could then go either into maintenance and repair in case the log analysis revealed any pending issues or into long-term storage. The re-commissioning phase puts a device into operation again, that is, validates its integrity and tests the device functionality, before mission-specific configurations are applied again. At the end of each device's life, the secure equipment disposal ensures that all data remnants are destroyed before a device is destroyed and recycled.

The Project SD4MSD and its Consortium

In order to attain these ambitious goals and finally ensure the wide applicability of the developed solutions and procedures, the project consortium consists of a vital mix of a strong academic partner with deep knowledge in secure system design (Austrian Institute of Technology), a vendor of physical tablets with rich experience in implementing these solutions (MUSE Electronics GmbH) and the Austrian Ministry of Defence as end-user partner. SD4MSD is a 30-month national research project running from 2021 to 2023 and is funded by the Austrian defence research programme FORTE of the Federal Ministry of Finance (BMF)

References:

[1] MIL-STD-810H: Environmental Engineering Considerations and Laboratory Tests. U.S. Department of Defense, 2019.

[2] S. Sultan, I. Ahmad, T. Dimitriou, "Container security: issues, challenges, and the road ahead," IEEE access, vol. 7, 52976-52996, 2019.

[3] L. F. Rahman, T. Ozcelebi and J. Lukkien, "Understanding IoT systems: a life cycle approach," Procedia Computer Science, vol. 130, 1057-1062, 2018

Please contact:

Florian Skopik, AIT Austrian Institute of Technology, Austria
florian.skopik@ait.ac.at

SPACE: Scalable Parallel Astrophysical Codes for Exascale

by Marisa Zanotti (EnginSoft), Andrea Mignone (University of Torino) and Manolis Marazakis (ICS-FORTH)

In astrophysics and cosmology (A&C), high-performance computing (HPC)-based numerical simulations are invaluable instruments to support scientific discovery. The efficient and effective exploitation of exascale (and beyond) computing capabilities will be key to paving the way for scientific discovery in this scientific domain, but requires a coordinated effort towards adapting A&C codes for space applications on exascale HPC systems.

In astrophysics and cosmology (A&C), HPC (high-performance computing)-based numerical simulations are invaluable instruments to support scientific discovery. They represent essential tools for modelling, interpreting, and understanding the complex physical processes behind the observed sky. Advances in computational power promise a wealth of ground-breaking scientific discoveries by making ever greater numerical simulations feasible, provided that equally advanced tools are created to exploit these computational resources. Moreover, the outstanding quality and volume of observational data generated by the current and next generation of instruments (eg. LOFAR, MeerKAT, MWA, EUCLID, SKA) poses exceptional challenges to their theoretical interpretation and will require novel theoretical and numerical simulation capabilities (codes, algorithms and tools) able to investigate the physical processes behind the observed phenomena with unprecedented quality, resolution and reliability, allowing their interpretation and paving the way for scientific discovery.

For a field like A&C, which is generally unsuited to direct experiments, the significance of this development cannot be underestimated. Future exascale computing systems are expected to have extremely complex, heterogeneous architectures [1]. The currently used numerical simulation codes are not suitable for use on such systems, since they were not purposely designed for them and therefore cannot effectively take advantage of the superior processing capabilities expected. Therefore, the efficient and effective exploitation of exascale (and beyond) computing capabilities will be key to achieving the anticipated exceptional results in A&C, combined with high-performance data processing, interactive real-time analysis, and data visualisation to explore large volumes of data, compare observations with simulations, and finally implement multi-wavelength and multi-messenger science. Nowadays, a limited number of numerical applications, several of which are developed and maintained in Europe, represent the state-of-the-art in A&C simulations. However, although they are production codes used in cutting-edge simulations, they also require a substantial effort to evolve their computational paradigms from the petascale to the exascale era.

The main goal of the SPACE Centre of Excellence is to enable current astrophysical and cosmological codes to be used on the pre-exascale HPC architectures funded by the EuroHPC JU

and made available at the end of 2022, as well as on upcoming exascale architectures, by re-designing or adapting the existing computational tools for next-generation HPC hardware platforms. SPACE brings together scientists, community code developers, HPC experts, hardware manufacturers and software developers in co-design activities (Figure 1) to re-engineer eight of the most widely used European A&C HPC codes into new products that can efficiently exploit future computing architectures.

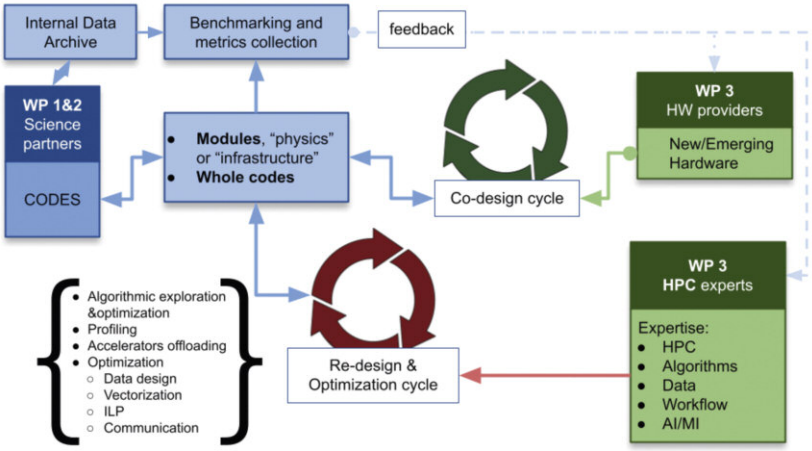


Figure 1: Re-design/optimisation and co-design cycles.

These eight A&C HPC codes represent 70% of the HPC A&C simulations, and were selected after an extensive analysis of their features and capabilities. They will initially be prepared to adequately exploit the pre-exascale systems with a view to their transition to exascale systems and beyond. At the same time, SPACE will work to advance workflows and data processing based on machine learning and visualisation applications, and to enhance their scalability on exascale systems. Contrasting the results from numerical simulations with the torrent of complex observational data from the new generations of ground- and space-based observatories will be the fundamental method to provide new insights into astronomical phenomena, the formation and evolution of the universe, and the fundamental laws of physics.

SPACE brings together a significant fraction of the numerical A&C European community from eight different countries, having as partners: University of Turin (Coordinator), INAF, CINECA, E4 and EnginSoft Spa from Italy; Universiteit Leuven from Belgium; VSB-Technical University of Ostrava from Czechia; CNRS and BULL from France; LMU and Goethe University from Germany; FORTH from Greece; University of Oslo from Norway; BSC from Spain.

Furthermore, SPACE will promote the adoption of general and community standards for data products based on FAIR principles [2], and the interoperability of data and applications based on the technology standards and best practices of the International Virtual Observatory Alliance (IVOA) [3].

SPACE will also implement the selected applications and foster their use by means of an outreach and training programme aimed at creating a broad and skilled talent pool in Europe to boost the use of high-performance and high-throughput solutions in academia in order to pave the way for the transition to exascale technologies and beyond. More information on the work plan and outcomes of SPACE is available via the project's website [L1] and social media channels [L2, L3].

SPACE CoE is funded by the European Union. It has received funding from the European High Performance Computing Joint Undertaking and from Belgium, the Czech Republic, France, Germany, Greece, Italy, Norway, and Spain under grant agreement No. 101093441.

Links:

[L1] https://www.space-coe.eu/
[L2] https://kwz.me/hxc
[L3] https://twitter.com/CoeSpace50804

References:

[1] M. Schulte, et al., "Achieving exascale capabilities through heterogeneous computing," IEEE Micro, vol. 35, no. 04, pp. 26–36, 2015.
[2] M. Wilkinson, M. Dumontier, I. Aalbersberg, et al., "The FAIR guiding principles for scientific data management and stewardship," Scientific Data, 3, 160018, 2016, https://doi.org/10.1038/sdata.2016.18
[3] G. B. Berriman, The IVOA Executive Committee, The IVOA Technical Coordination Group and The IVOA Community, Proc. of The International Virtual Observatory Alliance in 2020, 2020. https://doi.org/10.48550/arXiv.2012.05988

Please contact:

Maria Zanotti, EnginSoft Spa, Italy
m.zanotti@enginsoft.com

Call for Participation

Denotational Engineering

by Andrzej Blikle (Institute of Computer Science, Polish Academy of Sciences)

This project is devoted to the development of programming languages in a way that guarantees the existence of denotational semantics for them. Once such a language has been developed, we can derive such program-construction rules, which guarantee the correctness of programs constructed by their means.

It is a well-known fact that every user of a software application has to accept a disclaimer, such as, "There is no warranty for the program to the extent permitted by applicable law. Except when otherwise stated in writing, the copyright holders and/or other parties provide the program 'as is' without warranty of any kind … ".

I believe that the cause of this situation is a lack of mathematical tools for software engineers (commonly available for all other engineers) that would guarantee the functional reliability of programs based on how they have been designed and devel-

oped. Even though a lot of research has addressed this problem since 1960, and several mathematical methods have been developed, none has become a standard in the software industry.

The project is devoted to the elaboration of two methods. The first concerns the development of programming languages along with their denotational models (semantics) and is called "denotational engineering". The second is devoted to developing programs that are correct with respect to their specifications, the latter nested in the codes of programs.

In our approach, the design process of a programming language starts from the development of its algebra of denotations, called AlgDen, described in a metalanguage, MetaSoft. Once such a description is completed, we derive from it a corresponding syntax described by an equational grammar. The syntax also constitutes an algebra that we call AlgSyn. Our method guarantees the following:

1. There exists a (unique) denotational semantics of AlgSyn, which is a homomorphism – let's denote it by Sem – from AlgSyn into AlgDen.
2. A context-free grammar corresponding to AlgSyn may be half-algorithmically derived from the description (in MetaSoft) of AlgDen.
3. A denotational definition of Sem (in MetaSoft) may be algorithmically derived from the definitions of both algebras.
4. The definition of Sem may be transformed into an interpreter of the designed language. Again, this task may be performed, to a large extent, algorithmically.

Once we have a language with full denotational semantics, we can derive a repertoire of program constructors that guarantee the correctness of constructed programs. In this approach, each program consists of a "programming layer", which is a code, and a "descriptive layer", which constitutes this program's specification written in a formalised dialect of MetaSoft. A program is called "correct" if its programming layer is totally correct with clean termination (no error messages) [1] with respect to its descriptive layer. This approach offers an alternative to first developing programs and only then trying to prove them correct.

It is a well-known fact that the method of proving programs correct faces two significant challenges:

- Since a proof is usually longer than a theorem, a proof of program correctness must be longer than the program itself. That is, frankly speaking, somewhat discouraging.
- Even more discouraging is that programs whose correctness we intend to prove are practically never correct.

In our method of developing correct programs, rather than proving programs correct, we do not avoid proofs at all, but we simplify them to a large extent. Whenever we intend to apply a correctness-preserving constructor to some earlier-developed correct programs, we usually have to prove an implication between two formulas that are mathematically relatively simple but may contain many variables. There may be as many variables as declared in the program, which can make the corresponding proof hardly manageable by hand. In this case, however, we may easily use a theorem prover.

The project's current state has been described in a draft version of a book [L1] and in an article published by Springer [L2]. The book includes material taught twice as a facultative one-semester course at the Department of Mathematics, Informatics, and Mechanics of the University of Warsaw in the years 2020 and 2021. The first course ended with the development (by students) of an interpreter (written in OCaml) of a programming language, Lingua-WU (WU for Warsaw University), developed by the method of denotational engineering. The language is imperative, strongly typed, and includes recursive procedures with mutual recursion. The book also contains a list of sound program-construction rules, with appropriate logical background based on three-valued predicates, and a denotational model of a subset of Stuctured Query Language (SQL). Further research in progress tackles denotational models of:

- concurrency at the level of simple Petri nets
- object-oriented mechanisms
- a subset of OCaml to study its mechanism of parametric types.

At the mathematical level, denotational engineering is based on the following:

1. Fixed-point theory in partially ordered sets
2. A calculus of binary relations
3. Formal-language theory and equational grammars
4. Fixed-point domain equations based on so-called "naive denotational semantics" [3], where denotational domains are usual sets rather than Scott-Strachey reflexive domains
5. Many-sorted algebras
6. The development of a language from denotations to syntax [2]
7. Abstract errors as a tool for the description of error-handling mechanisms
8. Three-valued predicate calculi
9. A theory of total correctness of programs with clean termination [1].

The project is currently not associated with any institution and is developed by me with two colleagues from the University of Warsaw. The project offers several opportunities for MS and PhD dissertations in theory and software engineering as well as several research problems to be addressed [L3]. It is open to all interested parties.

Links:

[L1] https://moznainaczej.com.pl/what-has-been-done/the-book

[L2] https://moznainaczej.com.pl/what-has-been-done/on-the-development-of-a-denotational-model

[L3] https://moznainaczej.com.pl/what-remains-to-be-done

References:

[1] A. Blikle, "The Clean Termination of Iterative Programs," Acta Informatica, vol. 16, pp. 199-217, 1981.

[2] A. Blikle, "Denotational Engineering," Science of Computer Programming, vol. 12, North-Holland, 1989.

[3] A. Blikle A. Tarlecki, "Naïve denotational semantics," Information Processing 83, R.E.A.Mason, Ed, Elsevier Science Publishers BV (North-Holland), 1983.

Please contact:

Andrzej Blikle, Institute of Computer Science, Polish Academy of Sciences, Poland

Sponsored Contribution

TRAPEZE: Transforming Data Management for All

by Lauro Vanderborght (Digitaal Vlaanderen), Martin Kurze (Deutsche Telekom) and Ramon Martin de Pozuelo (CaixaBank)

The TRAPEZE project, titled "Transparency, Privacy, and Security for European Citizens," is making remarkable progress as it moves forward in its journey. With its objective of demonstrating the TRAPEZE prototype solution in real-world scenarios, the three project use cases are advancing splendidly. These use cases will exemplify how the TRAPEZE solution overcomes present-day limitations and revolutionizes the way enterprises, public administration, and citizens interact with their sensitive information.

Three real-world use cases led by Informatie Vlaanderen, Deutsche Telekom, and CaixaBank demonstrate the capabilities of TRAPEZE outcomes. The development and status of the use cases are highly promising. The solution is designed to be flexible, robust, scalable, and ethically compliant. Its potential extends far beyond the project's conclusion, as it is set to be adopted by a broad range of entities and citizens, positively transforming various scenarios.

Informatie Vlaanderen: "My Citizen Profile"

In recent years, there has been a growing emphasis on citizen-centricity and secure data sharing as important aspects of digital transformation. Digital Flanders is a government agency from the Flemish Government that recognizes this and is working on a new infrastructure to address these needs. Their aim is to create a secure and standardized way for citizens to reuse government data, with a focus on providing an excellent user experience.

To achieve their goal, Digital Flanders is leveraging Solid, a technology that was invented by Tim Berners-Lee, the creator of the World Wide Web, and researchers from UGent. Solid technology provides a platform that enables users to control their own data and choose how and with whom they share it, while ensuring the data remains secure and private. One of the main advantages of Solid is that it allows multiple organizations to make use of the same data, being stored in decentralized stores called Pods.

Digital Flanders is building on this technology to create a state-of-the-art data-sharing infrastructure. They are also leveraging existing developments from their popular MyCitizensProfile platform, which enables citizens to access their own data and manage their interactions with government services.

One of the first use cases for Digital Flanders' new infrastructure will be with Randstad, a large HR group. Randstad will use diploma data from Digital Flanders during their application process. Solid simplifies Randstad's process by offering a user interface for authentication and consent to access diploma data from the applicant's Solid Pod.. This collaboration serves as a practical demonstration of Digital Flanders' infrastructure capabilities.

Digital Flanders is also part of the TRAPEZE consortium, which has a goal of investigating and setting up a privacy platform that allows citizens to assess which third parties have consent to use their data and audit how their data has been used. This platform will build on the foundations of the Solid technology and aims to take consent management to the next level.

Digital Flanders aims to utilize the Solid project blueprint in collaboration with Randstad to assess TRAPEZE's potential for enhanced consent management and seamless integration into their existing infrastructure. This will enable enhanced control and security for citizens' personal data.

Deutsche Telekom: Tools & Applications for "Data sharing via APIs"

Deutsche Telekom's (DT) concern is to make language and privacy policies defined in the TRAPEZE language as well as tools available for legal and commercially useful exchange/sharing of telco-specific personal data. These tools can then also be marketed by T-Systems (DT's subsidiary for IT service provisioning) in the "Data Intelligence Hub (DIH)". Both contexts require an automated, GDPR-compliant mechanism for formulating, applying, and managing rules for data sharing. These are formulated in privacy policies.

DT is actively contributing to the CAMARA Telco Global API Alliance. In this context, APIs for sharing data – including personal data – are provided for 3rd parties to make use of functions, features and data provided by telco carriers. For Telcos, this is a unique opportunity to finally monetize some of the data they host. DT pays a lot of attention to not harming its excellent reputation in terms of privacy and security. Thus, customer consent is collected in advance, and agreed privacy policies are used as a means of consent management.

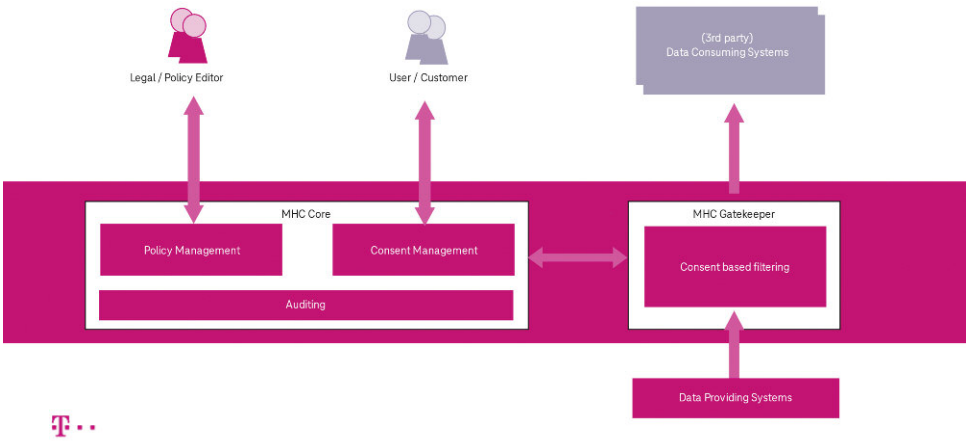


Figure 1: MHC architecture overview.

TRAPEZE language is used to define, share, manage and enforce consent (or rather "agreed privacy policies"). DT integrated TRAPEZE language, tools, and concepts in its' "Magenta Hyper Consent (MHC)" product. This product is targeted toward product owners and (in the CAMARA context) API monetization. Thus, there is no dedicated "TRAPEZE" user interface used, but rather DT/product-specific user interfaces are utilized to collect consent and to allow users to manage their privacy preferences.

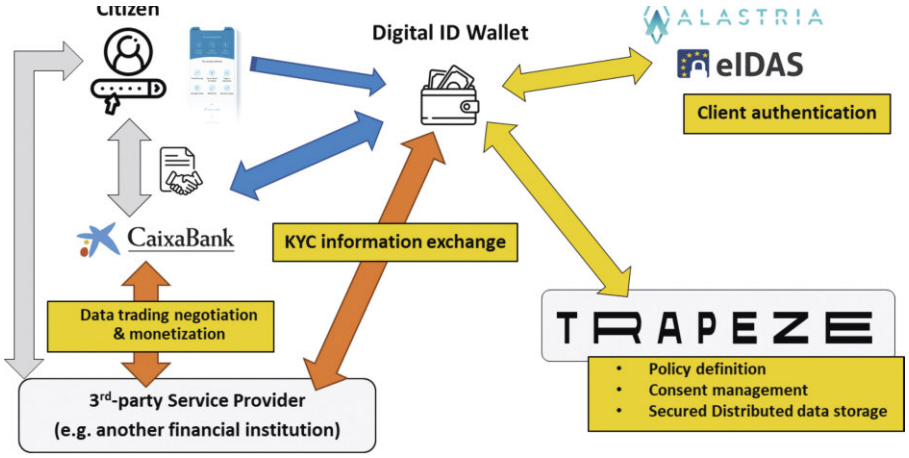


Figure 2: The Digital ID Wallet.

While the MHC Core deals with policy- and consent management (independently of actual data), the MHC Gatekeeper uses the policies to filter 3rd party data requests. Figure 1 shows the overall architecture from a technical point of view. Since MHC aims at B2B business and product managers, not directly at end customers, all components are built in a way to allow easy integration in new and existing products and services. It enables the DIH and other DT Business units to safely deal with personal data in the context of GDPR and other regulations.

A first application was trialed with DT's approach for consent management, the "group consent clause" which allows customers (i.e. citizens) to grant, revoke and manage their consent for data using and sharing. A key requirement is the open exchange format of privacy policies as it was developed in TRAPEZE and its predecessors.

CaixaBank: "Customers' Digital ID wallet"

Caixa Bank (CXB) wants to develop a "Customer ID Wallet" that allows the bank direct and transparent communication with clients about the usage of their data. It will be designed to enforce GDPR compliance and increase the data privacy security awareness of their clients as well as incorporate the bank's business requirements. No existing unified platform yet ensures security, privacy control, transparency, and trust among stakeholders. CXB wants to harness the potential of the TRAPEZE platform and its building blocks to address this challenge effectively.

Moreover, the European Commission's recent release on the European Digital Identity framework strengthens CXB's innovation perspective on trusted self-management of individual identity and data. This should help to streamline secure onboarding processes for new digital financial services, enhance overall security awareness, promote data privacy consciousness, and ultimately reduce successful social engineering attacks and impersonations. In that line, the Customers' ID Wallet pilot aims at developing an identity wallet that can work as a technical reference or complement the future EU Digital wallet, considering the digital identity verification means provided by the EU and Member States (when available) or any other trusted entity that works as an identity provider.

The pilot supports a key use case: enabling secure exchange of Know Your Customer (KYC) information between entities, ensuring banks collect and maintain up-to-date client information. This is essential due to Anti-Money Laundering (AML) regulations, which mandate the collection and maintenance of client information for all financial institutions. However, properly collecting, updating, and verifying the accuracy of this information from all clients poses a significant time-consuming challenge for both banks and citizens seeking their services. Presently, whenever a citizen intends to open an account with a new bank, they are required to provide the necessary personal and financial details.

What if we could collect and validate KYC information just once? This would simplify the process for both banks and citizens, and that's precisely what the "Customers' Digital ID Wallet" pilot aims to achieve. With this pilot, citizens can provide their information once to a single financial institution. The entity will validate the information as usual, but the Digital ID Wallet will securely retain and enable sharing of this attested information when the citizen seeks financial services from another bank. For this to happen, the customers must also be able to assess both the risks and the potential benefits of such actions (e.g. control with which entity they are sharing the data in order to identify them and their profile faster). The TRAPEZE platform will provide an easy and user-friendly way in which citizens can manage their data privacy policies and also review which entity has the consent to access which sensitive data from them and for which purpose.

As a result, Customers' Digital ID Wallet can improve the citizens' overall awareness of their data security and privacy risks, making them active players in the protection of their own data and finances.

Link: https://trapeze-project.eu/

Please contact:
Lauro Vanderborght, Digitaal Vlaanderen, Belgium
lauro@digita.ai

Martin Kurze, Deutsche Telekom, Germany
Martin.Kurze@telekom.de

Ramon Martin de Pozuelo, CaixaBank, Spain
rmartindepozuelo@caixabank.com

Call for Contributions

MISDOOM - The 5th Symposium on Multidisciplinary International Symposium on Disinformation in Open Online Media

Amsterdam, The Netherlands
21-22 November 2023

The Multidisciplinary International Symposium on Disinformation in Open Online Media (MISDOOM) is returning for its 5th edition on 21 and 22 November 2023. This time, the conference will be hosted by the National Research Center for Mathematics and Computer Science (CWI) at Amsterdam Science Park (Netherlands). MISDOOM values multidisciplinary research and is designed to be inclusive of different academic disciplines and practices.

The symposium provides a platform for researchers, industry professionals, and practitioners from various disciplines such as communication science, computer science, computational social science, political science, psychology, journalism, and media studies to come together and share their knowledge and insights on online disinformation.

Symposium Topics

Participants can discuss and contribute to the following list of topics:

- Cross-platform campaigns and their impact (e.g., diffusion of disinformation and manipulation, observations of campaigns and strategies, communication strategies, hate speech)
- Approaches to studying misinformation (e.g., qualitative approaches, case studies, quantitative approaches, experiments)
- User involvement with misinformation on various platforms (e.g., engagement, viewership)
- Counter-measures for mis- and disinformation and manipulation (e.g., censorship policies, behavioral changes, education, training, professional codices, legal actions)
- Factors contributing to misinformation beliefs or hampering corrections of false beliefs (e.g., political polarization, motivated reasoning, confirmation bias)
- Trending topics in mis- and disinformation research
- Automated fact-checking and misinformation detection
- Models for misinformation diffusion
- Human computation approaches for misinformation detection (crowdsourcing, human-machine interaction)
- Information quality (information quality dimensions, metrics, ethics of information quality)
- Generative AI tools and disinformation (e.g., ChatGPT, Midjourney, DALL-E) .

Important Dates

- Submission Deadline: 30 June 2023
- Notification: 28 August 2023
- Camera-ready: 11 September 2023
- Symposium: 21-22 November 2023

https://event.cwi.nl/misdoom-2023/submissions/

Call for Contributions

14th IFIP Trust Management Conference

The 14th IFIP International Conference on Trust Management (IFIPTM 2023) will be held in Amsterdam, The Netherlands on 18-20 October 2023.

The mission of the IFIPTM 2023 conference is to share research solutions to problems relating to trust and trust management in digital infrastructures. As society become increasingly digitalized, interactions are transferred to the digital domain, which introduces problems of establishing, maintaining and exploiting trust among entities and services in the digital domain, including related security and privacy issues. One indicator of the increasing importance of trust in the digital domain is the increase in legislation and regulation, such as the GDPR regulation and NIS 2 directive in Europe, but similar regulations are emerging around the World. The conference seeks to present solutions to all issues relating to trust and trust management in a digitalized society, and to identify new challenges and directions for future research.

IFIPTM 2023 will present novel research on all topics related to trust, security and privacy, including but not limited to those listed below:

- Trust in Information Technology
- Trust and Identity Management
- Socio-Technical and Sociological Trust
- Emerging Technology for Trust

Organizers

General co-chairs:

- Davide Ceolin, CWI, Amsterdam, The Netherlands
- Ehud Gudes, Ben-Gurion University, Israel
- Nurit Gal-Oz, Sapir Academic College, Israel

Program co-chairs:

- Tim Muller, University of Nottingham, United Kingdom
- Carmen Fernandez-Gago, University of Malaga, Spain

Important Dates

- Submission deadline:15 July 2023
- Author notification: 1 August 2023
- Conference: 18 - 20 October 2023

https://event.cwi.nl/ifiptm2023/

ERCIM "Alain Bensoussan" Fellowship Programme

The ERCIM postdoctoral Fellowship Programme has been established as one of the premier activities of ERCIM. The programme is open to young researchers from all over the world. It focuses on a broad range of fields in Computer Science and Applied Mathematics.

> "The ERCIM fellowship program motivated me to explore new research fields, meet experts and researchers from diverse backgrounds in order to get a broader perspective and bring new ideas to solve complex problems. Finally, one of the great learning from this program is, how to bridge the gap between academic research and tangible research outcomes like real time applications and products.
>
> **Shrutika SAWANT**
> Former ERCIM Fellow

The fellowship scheme also helps young scientists to improve their knowledge of European research structures and networks and to gain more insight into the working conditions of leading European research institutions. The fellowships are of 12 months duration (with a possible extension), spent in one of the ERCIM member institutes. Fellows can apply for second year in a different institute.

Where are the fellows hosted?

Only ERCIM members can host fellows. When an ERCIM member is a consortium the hosting institute might be any of the consortium's members. When an ERCIM Member is a funding organisation, the hosting institute might be any of their affiliates. Fellowships are proposed according to the needs of the member institutes and the available funding.

The fellows are appointed either by a stipend (an agreement for a research training programme) or a working contract. The type of contract and the monthly allowance/salary depends on the hosting institute.

ERCIM encourages both researchers from academic institutions and scientists working in industry to apply.

Why to apply for an ERCIM Fellowship?

The Fellowship Programme enables bright young scientists from all over the world to work on a challenging problem as fellows of leading European research centers. In addition, an ERCIM fellowship helps widen and intensify the network of personal relations and understanding among scientists. The programme offers the opportunity to ERCIM fellows:

- to work with internationally recognized experts,
- to improve their knowledge about European research structures and networks,
- to become familiarized with working conditions in leading European research centres,
- to promote cross-fertilization and cooperation, through the fellowships, between research groups working in similar areas in different laboratories.

Equal Opportunities

ERCIM is committed to ensuring equal opportunities and promoting diversity. People seeking fellowship within the ERCIM consortium are not discriminated against because race, color, religion, gender, national origin, age, marital status or disability.

Conditions

Candidates must:

- have obtained a PhD degree during the last eight years (prior to the application year deadline) or be in the last year of the thesis work with an outstanding academic record. Before starting the grant, a proof of the PhD degree will be requested;
- be fluent in English.

Application deadlines

Deadlines for applications are currently 30 April and 30 September each year.

Since its inception in 1991, over 790 fellows have passed through the programme. In 2022, 40 young scientists commenced an ERCIM PhD fellowship and 69 fellows have been hosted during the year. The Fellowship Programme is named in honour of Alain Bensoussan, former president of Inria, one of the three ERCIM founding institutes.

http://fellowship.ercim.eu

> "The ERCIM Fellowship is an excellent program and I recommend it for young researchers who have finished their PhD and are looking for opportunities to gain experience in European research institutions. It has a very simple and straightforward application process and the administration staff was always very helpful and supportive.
>
> **Léo FRANÇOSO DAL PICCOL SOTTO**
> Former ERCIM Fellow

Open Call for Innovative Extended Reality Tools and Applications Targeting Training and Educational Scenarios

A total funding of €2,100,000 will be distributed to projects for the development of novel XR-based applications primarily targeting the manufacturing domain. The total amount of funding per successful project will be in the range of €150,000 to €300,000. Each project may have up to three members including industrial partners, universities and research centers.

The XR2Learn project funded under the Horizon Europe framework Cluster 4-2022-HUMAN-01 is announcing its 1st Open Call.

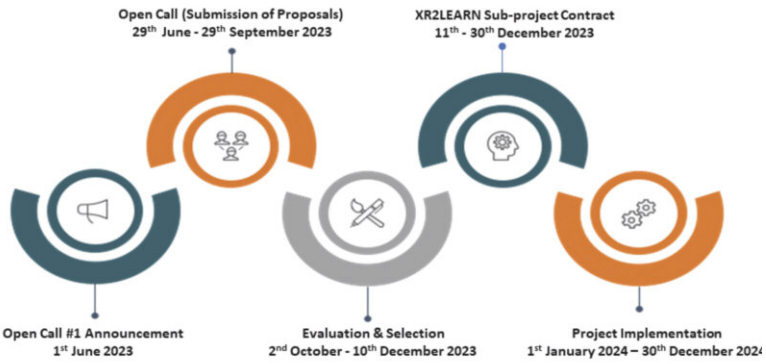
The objective of the XR2Learn – Open Call #1 is to support the adoption of Virtual, Augmented, and Mixed Reality (VR/AR/MR) in educational and training scenarios in industry 5.0 technologies as well as other educational sectors (e.g. healthcare and medical, manufacturing, construction and engineering).

An application may be submitted by an individual SME or a small consortium of up to three members in total to develop an immersive, interactive experience. In case of a consortium, the Project Leader/ Coordinator must be an SME, while the inclusion in the consortium as a member of a potential user/ early adopter (e.g. University, Academic or Training organization) may be a plus (although not mandatory).

The total funding distributed at Open Call #1 is €2,100,000. The total amount of funding per successful sub-project will be in the range of €150,000 to €300,000. Each industrial entity may receive between €60,000 and €200,000, while the total funding for all industrial partners should be at least 60% of the total funding. All partners are funded at 100%, through lump sum. 7-14 applications will be selected via this open call to enter the process, extend XR2Learn scope and deliver novel XR-based applications primarily targeting the manufacturing domain.

Submission

Submission will be enabled on Thursday, June, 29th, 2023, and will end on Friday, September 29th, 2023, at 17:00CET time (Brussels time). Selected projects are expected to start on January 1st, 2024.



Links:
[L1] https://xr2learn.eu
[L2] https://xr2learn.eu/open-calls/

Please contact:
Ioannis Chatzigiannakis
Sapienza University of Rome
and CNIT, Italy
ichatz@diag.uniroma1.it

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Call for Proposals

Dagstuhl Seminars and Perspectives Workshops

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is accepting proposals for scientific seminars/workshops in all areas of computer science, in particular also in connection with other fields.

If accepted, the event will be hosted in the seclusion of Dagstuhl's well known, own, dedicated facilities in Wadern on the western fringe of Germany. Moreover, the Dagstuhl office will assume most of the organisational/ administrative work, and the Dagstuhl scientific staff will support the organizers in preparing, running, and documenting the event. Thanks to subsidies the costs are very low for participants.

Dagstuhl events are typically proposed by a group of three to four outstanding researchers of different affiliations. This organizer team should represent a range of research communities and reflect Dagstuhl's international orientation. More information, in particular details about event form and setup, as well as the proposal form and the proposing process, can be found on

https://www.dagstuhl.de/dsproposal

Schloss Dagstuhl – Leibniz-Zentrum für Informatik is funded by the German federal and state government. It pursues a mission of furthering world class research in computer science by facilitating communication and interaction between researchers.

Important Dates

- *Next submission period:*
  October 15 to November 1, 2023
- *Seminar dates:*
  Between September 2024 and August 2025 (tentative).

TRAPEZE Project Webinar: Pioneering Privacy, Transparency, and Security for European Citizens

Online, 17 July 2023

The European TRAPEZE project is delighted to invite you to a special webinar where the project will be showcasing the remarkable achievements. This groundbreaking initiative has made significant advancements in safeguarding individual privacy while ensuring data transparency and security.

Join us online on 17 July 2023 from 9:30 to 13:00 CEST, as we unveil the outcomes of the TRAPEZE project, bringing together experts to present two compelling use cases and highlight cutting-edge privacy-preserving technologies developed during this extensive research endeavor.

The webinar will be divided into two parts:

Part 1: Use Cases by Caixa Bank and Deutsche Telekom - During this session, representatives from Caixa Bank and Deutsche Telekom will share their real-world experiences and demonstrate how the TRAPEZE project has positively impacted their organizations. Gain valuable insights into the practical implementation of privacy, transparency, sovereignty, and security principles within the banking and telecommunications sectors.

Part 2: Privacy Preserving Technologies In the second part of the webinar, we will delve into the innovative privacy-preserving technologies that have been developed as part of the TRAPEZE project. Our experts will provide a comprehensive overview of these advanced solutions, showcasing their potential to transform data handling practices across various industries while ensuring individual privacy remains intact.

By attending this webinar, you will:

- Discover tangible use cases demonstrating the TRAPEZE project's impact on Caixa Bank and Deutsche Telekom.
- Gain insights into state-of-the-art privacy-preserving technologies, revolutionizing data management practices.
- Learn how the TRAPEZE project contributes to privacy, transparency, sovereignty, and security for European citizens.
- Engage in a stimulating discussion with industry experts and thought leaders shaping the future of data protection.

Don't miss this unique opportunity to be at the forefront of privacy innovation in Europe. Register now and secure your spot for the TRAPEZE project webinar.

We look forward to welcoming you to this insightful webinar and fostering meaningful discussions about the future of privacy and data security.

Agenda and free registration at:
https://www.ercim.eu/events/trapeze-webinar

Prestigious Gödel Prize for Ronald de Wolf

Ronald de Wolf from CWI and his co-authors receive the 2023 Gödel Prize for outstanding papers in theoretical computer science.

Ronald de Wolf (CWI, UvA, QuSoft) and his co-authors receive the prestigious Gödel Prize for outstanding papers in theoretical computer science. The Gödel Prize is jointly awarded by the ACM Special Interest Group on Algorithms and Computation Theory (ACM SIGACT) and the European Association for Theoretical Computer Science (EATCS). The prize will be awarded during STOC 2023, one of the most important conferences in theoretical computer science, which takes place on 20-23 June 2023 in Orlando, Florida. This year, there are two winning articles. The other paper receiving the 2023 Gödel Prize is by Thomas Rothvoss.

Ronald de Wolf says: "I am very proud and humbled to win this prize along with my co-authors, and to be listed among the amazing papers and amazing researchers that have received this prize before". Earlier winners of the Gödel Prize include well-known researchers like Cynthia Dwork, Shafi Goldwasser, Johan Håstad, László Lovász, Peter Shor, Dan Spielman, Mario Szegedy and Avi Wigderson.

Travelling Salesman Problem

Authors Samuel Fiorini, Serge Massar, Sebastian Pokutta, Hans Raj Tiwary and Ronald de Wolf were given the award for their article 'Exponential Lower Bounds for Polytopes in Combinatorial Optimization'. One of its main conclusions was that a particular attempt to solve the famous travelling salesman problem cannot possibly work. Ronald de Wolf explains: "This paper refutes an attempt to solve hard computational problems such as Travelling Salesman (TSP). We know how to solve so-called linear programs efficiently, so since the 1980s researchers have been trying to write down a small linear program for TSP. If successful, this approach would have momentous consequences for efficient algorithms. However, our paper - which generalizes work by Yannakakis from 1988 - definitively showed that the approach is doomed to fail, by proving that every linear program that describes TSP needs to be exponentially large. The proof combines geometry, combinatorics, and even a connection with quantum communication theory."

At STOC 2012, Ronald de Wolf and the rest of the team already received a Best Paper Award for their work, and in 2022 they won the ACM STOC 10-year Test of Time Award. Ronald de Wolf won the ERCIM Cor Baayen Award in 2003.

https://www.cwi.nl/en/news/goedel-prize-for-ronald-de-wolf/
https://www.sigact.org/prizes/gödel.html

ERCIM
European Research Consortium for Informatics and Mathematics

ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.

ERCIM is the European Host of the World Wide Web Consortium.

CNR
Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1, 56124 Pisa, Italy
www.iit.cnr.it

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

CWI
Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
www.cwi.nl

RISE
RISE SICS
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

Fonds National de la Recherche Luxembourg
Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
www.fnr.lu

SBA Research
SBA Research gGmbH
Floragasse 7, 1040 Wien, Austria
www.sba-research.org/

simula
SIMULA
PO Box 134
1325 Lysaker, Norway
www.simula.no

FORTH
INSTITUTE OF COMPUTER SCIENCE
Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
www.ics.forth.gr

SZTAKI
Eötvös Loránd Research Network
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
www.sztaki.hu/

Fraunhofer
IUK-TECHNOLOGIE
Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
www.iuk.fraunhofer.de

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
www.cs.ucy.ac.cy/

inesc
INESC
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal
www.inesc.pt

UNIVERSIDAD DE MÁLAGA
Institute for Software Engineering and Software Technology
"Jose María Troya Linero", University of Malaga
Calle Arquitecto Francisco Peñalosa, 18, 29010 Málaga
https://gp.uma.es/itis

Inria
Institut National de Recherche en Informatique et en Automatique
B.P. 105, F-78153 Le Chesnay, France
www.inria.fr

Universty of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
www.mimuw.edu.pl/

i.S.I.
Industrial Systems Institute
I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
www.isi.gr

VTT
VTT Technical Research Centre of Finland Ltd
PO Box 1000
FIN-02044 VTT, Finland
www.vttresearch.com