relAI

BUILDING TRUSTWORTHY AI FOR A SAFER TOMORROW

relAI [rɪ'laɪ]
Konrad Zuse School
of Excellence in Reliable AI
2022-2024

»As AI systems become
more integrated into our lives,
it is imperative that these
systems function accurately and
comply with ethical values.
The Konrad Zuse School of Excellence
in Reliable AI marks a significant milestone
in advancing AI and ensuring
its reliability and trustworthiness.«

Markus Blume, Bavarian Minister for Science and the Arts

© StMWK/Böttcher

As AI systems become more integrated into our lives, it is imperative that these systems function accurately and comply with ethical values. The Konrad Zuse School of Excellence in Reliable AI marks a significant milestone in advancing AI and ensuring its reliability and trustworthiness.

In the face of the current global shortage of skilled professionals, the Zuse School of Excellence in Reliable AI stands as a beacon of hope and opportunity. We are addressing this shortage head-on by providing world-class education as we are equipping the next generation of AI experts with the knowledge and skills they need to become the AI experts of tomorrow. This is a long-term investment in our collective future! The Zuse School enhances Bavaria's appeal as a destination for international AI talents, positioning it as a global hub for AI innovation. This influx of talent enriches our academic and research environments and drives economic growth and technological advancement. The unique collaboration between our universities of excellence, Technische Universität München und Ludwig-Maximilians-Universität München, allows the Zuse School to use the strengths of both institutions to create a world-class educational environment. Together we are building an international lighthouse in reliable AI, providing unparalleled visibility for Bavaria. Focusing on mathematical and algorithmic foundations, medicine and healthcare, robotics and interacting systems as well as algorithmic decision-making, the Zuse School addresses key areas that are critical to the future of AI, setting new standards for reliability. This is in line with the Bavarian Hightech Agenda's goal to strengthen Bavaria's capabilities in key technologies. The Zuse School creates a strong community spirit among its members – it kind of feels like a relAI family. This special environment brings together students, researchers and fellows from academia and industry to collaborate and innovate in the field of reliable AI.

I would like to thank all partners and supporters of Zuse School, particularly DAAD and BMBF, the industry partners and international academic collaborators. Their great support is the key to making our vision a reality. I am confident that the Konrad Zuse School of Excellence in Reliable AI will be a catalyst for innovation, education and collaboration. Together we are shaping a future in which reliable AI improves the quality of life for all of us.

Munich, June 2024
Markus Blume, MdL
Bavarian State Minister of Science and the Arts

Directors

Dear readers,

In summer 2022, we launched the Konrad Zuse School of Excellence in Reliable AI (relAI) as a joint project between the two Universities of Excellence, Technische Universität München (TUM) and Ludwig-Maximilians-Universität München (LMU). It is one of three Zuse Schools in Germany, each with a unique focus area.

As the name implies, relAI focuses on the challenges of making artificial intelligence reliable. AI is becoming a vital part of our everyday lives, and making it reliable and usage safe is one of the most pressing issues both in research and industry. relAI's purpose is to bring the most talented young researchers from all over the world to Germany, create an outstanding and inspiring environment to support their work, and foster the collaboration between science and industry to achieve this very goal of making AI reliable and safe.

Our report is designed to show the progress we have made in setting up the Konrad Zuse School relAI over the past two years. We are proud to show what we did and achieved, from numerous events to award-winning papers. We will introduce the people behind the school: management, students as well as fellows and partners from both academia and industry — everyone that we like to call the members of our relAI family.

We would like to extend our gratitude to the German government, acting through the Deutsche Akademische Auslandsdienst (DAAD) for funding our Konrad Zuse School and making all of this visionary work possible. We would also like to thank both Universities, TUM and LMU, for their continued support. We are very grateful for the trust placed in us. And we would also like to thank everyone involved in making the school a success, no matter in what role they are contributing to relAI.

Sincerely,

Prof. Dr. Stephan Günnemann (TUM)
Prof. Dr. Gitta Kutyniok (LMU)
Directors of the Konrad Zuse School in Reliable AI

RELIABLE AI FOR A BETTER FUTURE

The current technological revolution is largely driven by remarkable advances in artificial intelligence (AI). While the enormous potential of AI is widely acknowledged, concerns about its reliability remain a significant barrier to its broader adoption in industry and society. Ensuring safety, security, and privacy are essential prerequisites for AI applications in public interest domains — such as guaranteeing that robots do not pose a threat to human life, or protecting data confidentiality.

Our relAI family

relAI is dedicated to advancing research and innovation in AI, focusing on creating reliable and impactful AI solutions. To achieve its vision, relAI builds on the distinct perspectives and commitment of our key groups: MSc and doctoral researchers, fellows, industry, and academic partners.

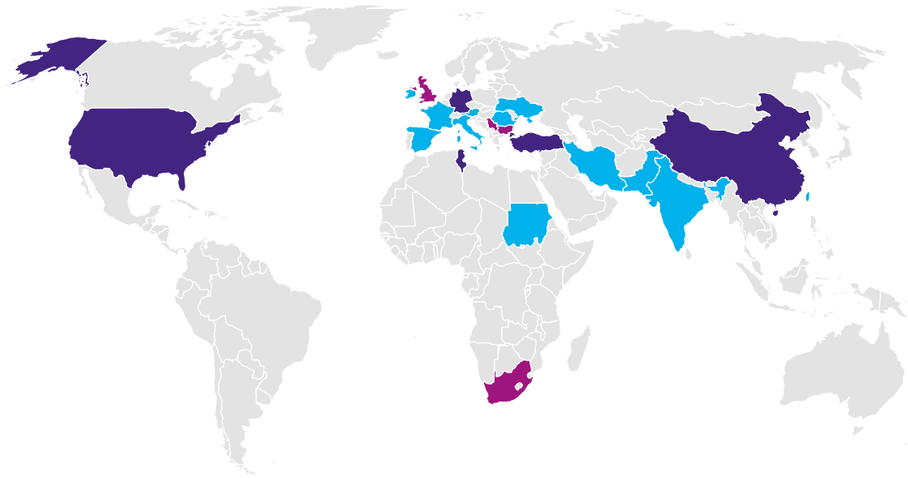
The Konrad Zuse School of Excellence in Reliable AI (relAI) envisions training future AI experts who uniquely blend technical proficiency with a deep understanding of AI reliability. Our groundbreaking and innovative AI program aims to educate top international candidates in the comprehensive development of reliable AI systems. This includes scientific knowledge, business acumen, and industrial experience, preparing them for roles in both industry and academia. Additionally, we conduct cutting-edge research to ready AI for deployment in critical applications.

In the first years of its existence, relAI has already succeeded in attracting exceptional talent from all over the world. Some of these individuals will get the chance to introduce themselves in this report. Also you will learn about the school itself, its people (students as well as fellows and partners), program, network and events.

STUDENTS

Our committed MSc and Doctoral students comprise the core of our school. They drive research in reliable AI and enrich the school through stimulating discussions, promoting reliable AI in outreach activities and broadening the relAI network through internships and academic exchanges.

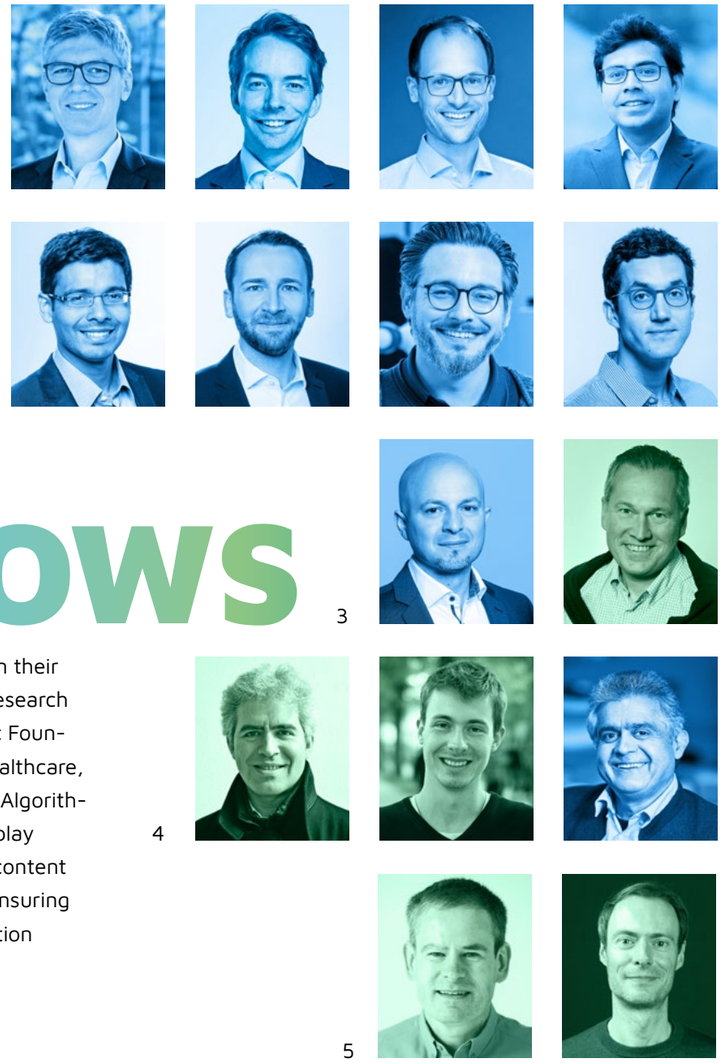
Since 2022, 23 MSc and 47 Doctoral students from around the world have joined the school through three different application calls. The first group of MSc students will graduate this year. Enabled by the consecutive relAI MSc & Doctoral program, some of them will continue in the program to pursue their Doctoral studies.



Origin of relAI students MSc & PhD PhD MSc

FELLOWS

Fellows in this program are experts in their fields, contributing to four primary research areas: Mathematical and Algorithmic Foundations of Reliable AI, Medicine & Healthcare, Robotics & Interacting Systems, and Algorithmic Decision-Making. These fellows play a crucial role in providing academic content and supervising Doctoral research, ensuring the development of the next generation of AI researchers.



1 **Albu-Schäffer, Alin** TUM
Althoff, Matthias TUM
Bauer, Stefan TUM, Helmholtz AI
Bhatotia, Pramod TUM
Bischl, Bernd LMU
Butz, Andreas LMU
Buyx, Alena TUM
Cremers, Daniel TUM
Drton, Mathias TUM
Eckert, Claudia TUM

2 **Feuerriegel, Stefan** LMU
Fortuin, Vincent TUM, Helmholtz AI
Ghoshdastidar, Debarghya TUM
Günnemann, Stephan TUM
Haddadin, Sami TUM
Heckel, Reinhard TUM
Hirche, Sandra TUM
Hüllermeier, Eyke LMU
Ingrisch, Michael LMU
Kaissis, Georg TUM
Kasneci, Enkelejda TUM

3 **Kasneci, Gjergji** TUM
Kauermann, Göran LMU
Kern, Christoph LMU
Kilbertus, Niki TUM, Helmholtz AI
Kreuter, Frauke LMU
Kuhn, Jochen LMU
Kutyniok, Gitta LMU
Lasser, Tobias TUM

4 **List, Christian** LMU
Maly, Johannes LMU
Navab, Nassir TUM
Ommer, Björn LMU
Rückert, Daniel TUM
Schmidt, Albrecht LMU
Schnabel, Julia TUM, Helmholtz Munich
Schöllig, Angela TUM
Schuller, Björn TUM

5 **Schütze, Hinrich** LMU
Sterkenburg, Tom LMU
Swikir, Abdalla TUM
Theis, Fabian TUM
Tresp, Volker LMU
Vieluf, Solveig LMU
Wachinger, Christian TUM
Zöller, Mark LMU

ACADEMIC PARTNERS

relAI cooperates with a number of academic Institutions all over the world. These collaborative relationships are intended to enhance the quality, scope and impact of our research endeavors, facilitate networking for our students and fellows, and allow for knowledge exchange and research stays.

EUROPE

- Data Science @ Uni Vienna
- Center for Intelligent Systems (CIS) at École Polytechnique Fédérale de Lausanne (EPFL)
- Center for Responsible AI Technologies coordinated by Munich School of Philosophy
- Helmholtz München
- Alan Turing Institute

ASIAN PACIFIC

- Center for AI Research (CAIRE) at Hong KongUniversity of Science and Technology (HKUST)

THE AMERICAS

- NSF AI Institute for Advance in Optimization at Georgia Institute of Technology (AI4OPT)
- Center for Data Science (CDS) at New York University (NYU)
- Center of Data Science and AI Research (CeDAR) at University of California, Davis (UC Davis)
- Center for Statistics and Machine Learning (CSML) at Princeton University
- Mathematical Institute for Data Science (MINDS) at John Hopkins University
- Oden Institute for Computational Engineering and Sciences at University of Texas at Austin
- Rhodes Information Initiative (RII) at Duke University
- Stanford Data Science Centre (SDS) at Stanford University

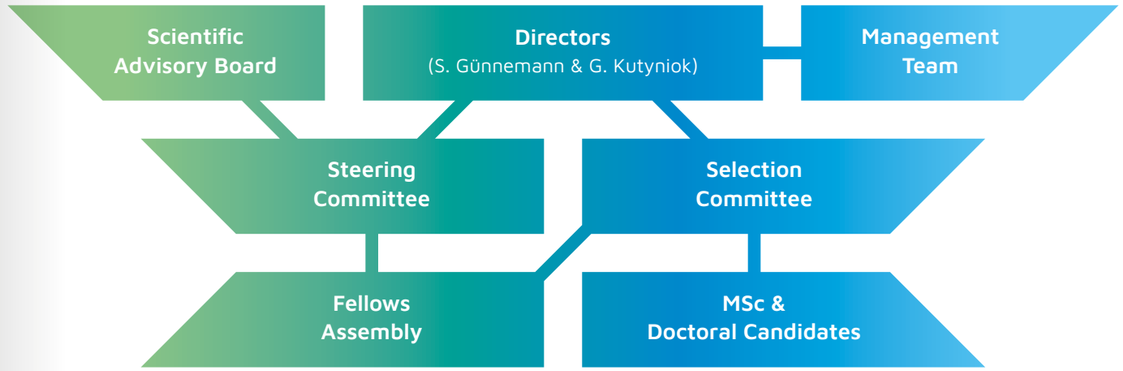
INDUSTRY PARTNERS

Cooperation with industry partners is another important pillar supporting the structure of our school. Part of relAI's curriculum are internships for MSc students, also Doctoral students have one industry member on their Transdisciplinary Thesis Advisory Committee (TTAC), so they receive feedback on their research and personal development from an application perspective. Many renowned companies have already joined in on a partnership with us, and the network is still growing.

ALLIANZ
BMW
BOSCH
CELONIS
DENSO
FRAUNHOFER INSTITUTE FOR...
...APPLIED AND INTEGRATED SECURITY AISEC
...COGNITIVE SYSTEMS IKS
...INTEGRATED CIRCUITS IIS
GOOGLE
IMFUSION
INFINEON
LINDE
MUNICH RE
SAP
SIEMENS
SIEMENS HEALTHINEERS
THYSSENKRUPP
UNTERNEHMERTUM
VOLKSWAGEN

Our Organizational Structure

relAI is structured to ensure effective governance and comprehensive oversight, encompassing several key bodies that each play a crucial role in its operation.



Directors: Prof. Dr. Stephan Günnemann (TUM) and Prof. Dr. Gitta Kutyniok (LMU) serve as directors of the Zuse School relAI.

Management Team: The Management Team supports the day-to-day operations of the organization. This office oversees administrative functions and coordinates activities across the school.

Steering Committee: The Steering Committee plays a pivotal role in guiding the school's projects and initiatives. Comprising senior members from various departments of both TUM and LMU, this committee ensures that projects align with the strategic goals. It is comprised of the directors plus four representatives from our four research areas from both universities.

Selection Committee: The Selection Committee is tasked with the critical responsibility of recruiting and selecting candidates for our Doctoral positions and MSc scholarships. It consists of all Steering Committee members plus four more fellows from both universities.

Scientific Advisory Board: The Scientific Advisory Board provides expert guidance on scientific matters, ensuring that the school's research and development efforts are of the highest quality. This board consists of distinguished international scientists who offer insights and recommendations on scientific projects and priorities.

Fellows Assembly: Once a year, all relAI Fellows meet to be informed about progress of the past year, to discuss future developments and elect new members for the Selection and Steering Committees.

Students: relAI students participate actively in the school's organization by taking responsibility for the relAI Blog, alumni network, and Wiki and planning excursions and talks.

EXPLORING NEW FRONTIERS

relAI's research mission is to train future generations of AI experts in Germany who combine technical brilliance with awareness of the importance of AI's reliability. As a central component of the Munich AI ecosystem, relAI is committed to making artificial intelligence more safe, secure, responsible and protective of individual privacy. It encourages responsible use to exploit AI's vast potential for the benefit of humanity, to advance insight into and debate on machine learning including its ethical and societal dimensions, and to strengthen "AI made in Germany."

»The scientific program of relAI contributes to the end-to-end development of reliable AI, covering different branches of applied research on the basis of profound mathematical and algorithmic foundations.«

Research projects are conducted by professors, post-docs, Doctoral candidates and MSc students in academic departments or research centers. Research dissemination is not only geared towards the academic world but also toward key innovative industries such as medicine and healthcare, robotics and interactive systems, as well as algorithmic decision-making.

The scientific program of relAI contributes to the **end-to-end development of reliable AI**, covering different branches of **applied research** on the basis of profound **mathematical and algorithmic foundations**. This theoretical grounding of AI applications is a distinguishing feature of relAI: Our conception of reliability involves the demand for a rigorous formal description of properties as well as provable guarantees. Only such guarantees will create the trust and confidence needed for practical adoption of AI, without reservations.

The research program combines Mathematical and Algorithmic Foundations of Reliable AI along with domain knowledge in three core application domains (as visualized in the figure on the right): Medicine & Healthcare, Robotics & Interacting Systems, and Algorithmic Decision-Making. For these applications, which are of major importance for Germany, reliable AI methods are most urgently needed. Thus, the school's research addresses a highly impactful and innovative topic with core societal demands in areas of public interest.

Each of the school's four research areas (in green) cover central themes of reliable AI (in blue):

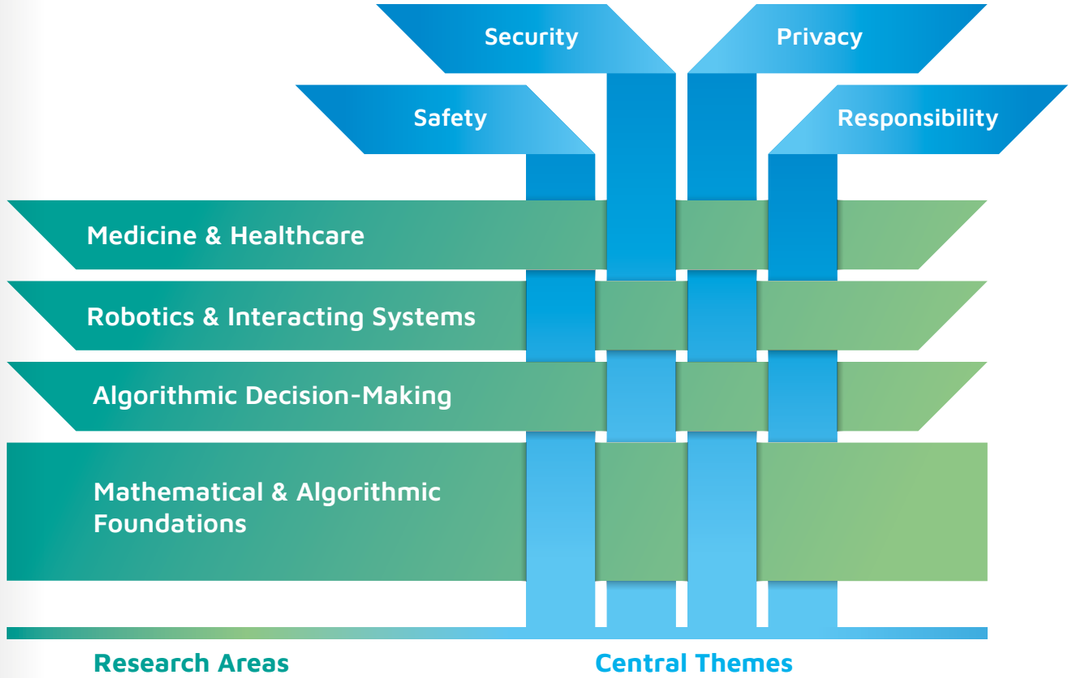
Safety, i.e., ensuring that AI systems (e.g. robots) do not cause any harm or danger.

Security, i.e., making AI systems resilient against threats, external attacks, and information leakage, e.g. avoiding manipulation of decision-making systems for criminal purposes.

Privacy, i.e., ensuring protection and confidentiality of (individual) data and information, such as medical AI systems incorporating sensitive patient data.

Responsibility, i.e., developing AI systems while taking societal norms, ethical principles, and human needs into consideration, for example, by explaining AI decisions and protecting individuals against discrimination.

By combining foundational AI research with core applications, we foster a strong interdisciplinary environment at relAI.



Since the start of relAI, the school has published findings in more than 70 scientific publications. The following examples illustrate the research taking place in relAI.

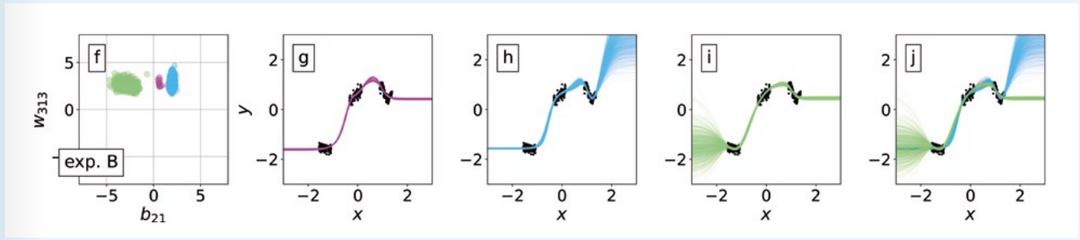
Mathematical & Algorithmic Foundations

Reliability of AI with all its facets can only be achieved through a profound understanding of its foundations. In fact, the gap between theory and practice of AI methodologies is one of the key obstacles for deriving comprehensive guarantees as required by critical applications. Supporting our goal of reliable AI, the general research challenges we address are twofold. Firstly, we work on establishing theoretical guarantees for AI. This includes expressivity of AI models, analysis of learning algorithms, generalization capabilities of trained AI systems, and aspects such as robustness, aiming predominantly at concrete error bounds and certification. A particular challenge is posed by novel and highly complex architectures, such as graph neural networks or transformers. Secondly, to support reliability, we research algorithmic foundations of AI on relevant topics, such as IT security, federated learning, distributed systems, and causal modeling, thereby ensuring a tight link to the application domains and their practical realization.

The paper "Towards efficient MCMC sampling in Bayesian neural networks by exploiting symmetry" is the result of a cooperation of multiple relAI Fellows and Doctoral students. It received the best paper award of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2023).

Wiese, J. G., Wimmer, L., Papamarkou, T., Bischl, B., Günnemann, S., & Rügamer, D. (2023). Towards efficient MCMC sampling in Bayesian neural networks by exploiting symmetry. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 459-474). Cham: Springer Nature Switzerland.

Bayesian inference in deep neural networks is difficult because of the complex, multi-dimensional nature of their parameter landscapes. Traditional methods like Markov chain Monte Carlo (MCMC) can accurately recover the true parameter distributions but are impractically slow and costly for modern, large-scale neural networks. An alternative approach involves local methods, which focus on specific parameter regions that are easier to manage mathematically. While these local methods can produce good results, they do not fully capture the complexity of the parameter landscape. The main achievements of this project are:



After symmetry removal, the remaining modes in the weight space represent distinct predictive functions that are highly relevant to uncertainty quantification

Exploiting Symmetries:
This research suggests that the trade-off between exact but expensive methods and faster but less accurate methods can be reduced by using the symmetries in the neural network parameter landscape. These symmetries occur because different parameter values can produce the same output due to neuron interchangeability and certain activation functions.

Theoretical Restriction:
The study shows that the posterior predictive density in Bayesian neural networks can be confined to a simplified set of parameters that are free from symmetry.

Efficient Sampling:
By establishing an upper limit on the number of Monte Carlo chains needed to represent the diverse functions of the neural network, the researchers propose a more efficient method for Bayesian inference.

Practical Implications:
Experimental results indicate that this method allows for efficient sampling, paving the way for more accurate uncertainty quantification in deep learning.

This research presents a promising approach to make Bayesian inference more feasible and accurate in deep neural networks by leveraging inherent symmetries, thus combining the strengths of both exact and approximate methods.

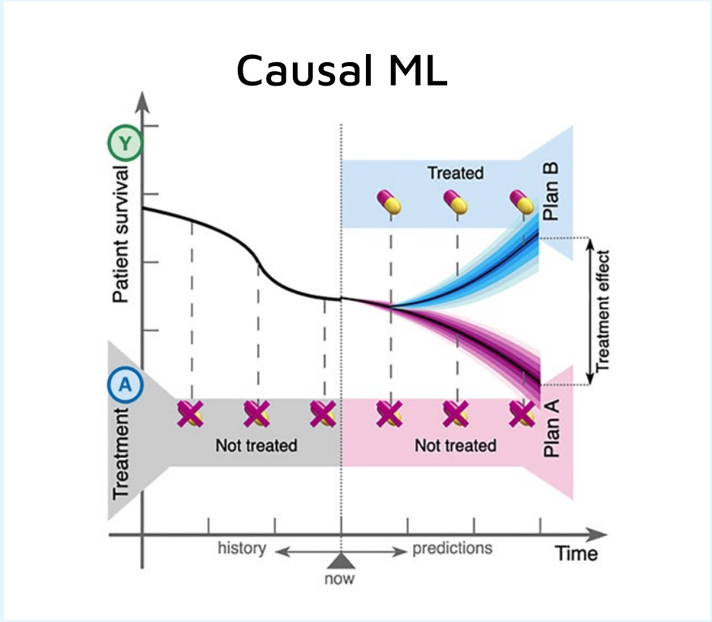
Medicine & Healthcare

AI has the potential to fundamentally transform the future of medicine and healthcare by enabling earlier and more accurate diagnosis and better treatment, leading to improved outcomes for patients and increased efficiency in healthcare. The emergence of AI for medicine and healthcare also offers a number of transformative opportunities for economic growth. Examples include prevention and early detection, e.g. AI for wearable devices as well as AI for screening (e.g. mammography). A key requirement for the successful deployment of AI in clinical environments is the development of safe, secure, and trustworthy ML techniques. In relAI, we are working on robust and data efficient learning, privacy preservation, and interpretable deep learning.

The paper "Causal machine learning for predicting treatment outcomes" is the result of cooperation between multiple relAI Fellows and Doctoral students.

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., & van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. Nature Medicine, 30(4), 958-968.

Causal machine learning is leveraged to predict treatment outcomes in medicine.



Causal machine learning (ML) offers advanced methods for predicting treatment outcomes, such as efficacy and toxicity, allowing for personalized clinical decision-making. It aims to predict changes in patient outcomes due to treatments, a causal quantity not addressed by traditional ML. In estimating individualized treatment effects, causal ML can use data from clinical trials and real-world sources like electronic health records. This research highlights the benefits of causal ML over traditional methods, outlines its key components, and provides recommendations for its reliable use in clinical settings. Main advances of this research are:

Personalized Predictions: Causal ML can predict individual patient outcomes, enabling personalized treatment plans based on specific patient profiles.

Handling Complex Data: It can manage high-dimensional and unstructured data, making it suitable for multimodal datasets, including images, text, and genetic information.

Estimating Heterogeneous Treatment Effects: It can identify variations in treatment effects across different patient subgroups, enhancing the precision of medical care.

To estimate treatment effects accurately, certain assumptions about the causal structure must be made. For example, one must account for how changes in treatment affect other patient characteristics. Traditional ML might ignore these dependencies, leading to incorrect predictions. To ensure that causal ML is embedded in a causal framework to address these issues, this research gives the following recommendations:

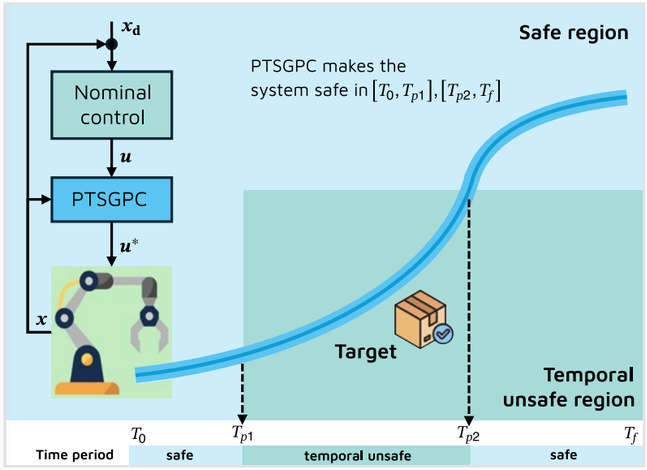
Check Assumptions:
Validate the assumptions underlying the causal ML model, such as ensuring there are no unmeasured confounders and that treatment of one patient does not affect another's outcome.

Model Validation:
Compare estimated treatment effects from real-world data with those from randomized controlled trials to validate the reliability of causal ML methods.

Transparency and Reporting:
Clearly state assumptions, chosen methods, and robustness checks. Use preregistered protocols to mitigate risks of false positives and selective reporting.

Robotics & Interacting Systems

Engineers and computer scientists are currently developing autonomous systems with AI techniques as a core component. This provides endless possibilities but also comes with enormous challenges regarding safety, security, and privacy. For example, how is it possible to guarantee safety of an autonomous agent (e.g. a robot in a human environment) under all circumstances, given that a designer cannot foresee all future situations? How can we balance the advantages of AI cloud computing with the increased risk of security violations? How can data be leveraged to adapt to the needs of a human user while bearing privacy concerns in mind? To answer such questions, relAI focuses on safe, secure, and privacy-preserving AI in the context of autonomous agents and interacting systems.



Trajectory for robot manipulator under PTSC and PTSCGP

The paper "Learning-Based Prescribed-Time Safety for Control of Unknown Systems with Control Barrier Functions" is the result of research by multiple relAI Doctoral students, supervised by a relAI Fellow.

Huang, T. Y., Zhang, S., Dai, X., Capone, A., Todorovski, V., Sosnowski, S., & Hirche, S. (2024). Learning-based prescribed-time safety for control of unknown systems with control barrier functions. IEEE Control Systems Letters.

Control systems often need to meet constraints for limited periods, such as during a robot-human handover. These temporary constraints allow flexibility to improve control performance. However, existing methods tend to enforce constant safety, leading to overly cautious behavior. Newer approaches focus on ensuring the system returns to safety within a finite time but these approaches struggle with unknown dynamics. The main advances of this project are:

New Method:
Prescribed-Time Safe Gaussian Process Control (PTSGPC) introduces a Gaussian process-based control method that combines backstepping and control barrier functions to ensure safety within specific time windows for systems with unknown dynamics.

Theoretical Guarantees:
Provides rigorous proofs that the system remains or returns to safety within a set time with high probability, regardless of initial conditions.

Effectiveness Demonstration:
Demonstrates the method's effectiveness through a simulation of a two-link robot manipulator.

PTSGPC is better at handling uncertainties, keeping the robot manipulator safe and close to the desired trajectory throughout the process. It reliably meets safety constraints with high probability, as demonstrated through various performance metrics and simulations.

In conclusion, this research resulted in a novel and safe learning control method for systems with control inputs, ensuring safety within a given time regardless of initial conditions. By integrating time-varying design and Gaussian process regression in barrier functions, the method guarantees high-probability safety, validated through theoretical and simulation results.

Algorithmic Decision-Making

Ever more applications in AI consider prescriptive modeling in the sense of learning a model that stipulates appropriate decisions or actions to be taken in real-world scenarios: Which medical therapy should be applied? Should this person be hired for the job? These sorts of decisions are increasingly automated and made by algorithms instead of humans, often relying on AI methods. The work on AI-based methodologies for reliable algorithmic decision-making (ADM) implies the need to address specific technical issues, such as the lack of an objective "foundational truth" underlying every prediction, and learning from partial training information, comprising feedback about the decision made, while lacking information about counterfactuals. Methodological research on ADM is complemented by more application-oriented research on reliable decisions in business and management.

The paper "One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions" was the result of research by a relAI Doctoral student, supervised by a relAI Fellow.

Simson, J., Pfisterer, F., & Kern, C. (2024). One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1305-1320).

Algorithmic decision making (ADM) is increasingly used to automate decisions that were traditionally made by humans. The impact of ADM systems heavily relies on the choices made during their design, implementation, and evaluation. These choices can either mitigate or reinforce biases present in the data. Often, these decisions are made implicitly, without fully understanding their influence on the final system.

To address this lack of transparency, this work introduced a method called multiverse analysis, drawing on insights from psychology, to enhance algorithmic fairness. The method makes implicit decisions explicit, allowing to demonstrate their impact on fairness. By combining these decisions, the method creates a grid of all possible decision combinations, or "universes."

For each universe, the method calculates metrics of fairness and performance. This allows to examine how different decisions affect fairness and to assess the variability and robustness of fairness scores. The main advances this project offers are:

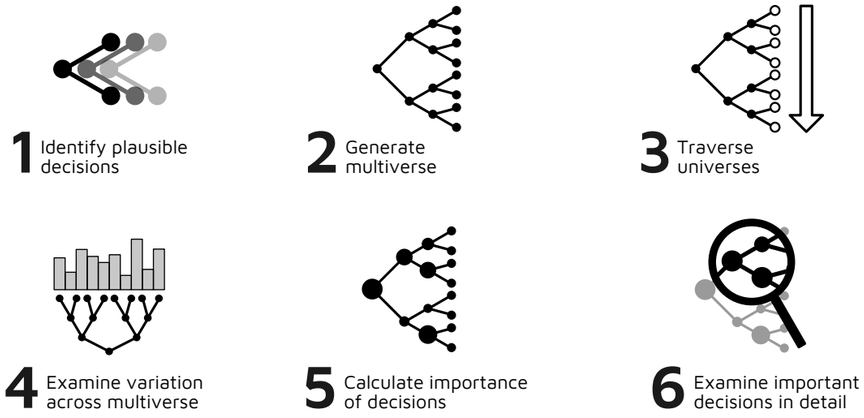
Explicit Decision-Making:
The research turns implicit design and evaluation decisions into explicit ones, providing clarity on how each decision impacts fairness.

Comprehensive Analysis:
By creating a multiverse of decision combinations, it is possible to investigate the full range of possible outcomes and their implications for fairness.

Robust Fairness Evaluations:
The method helps identify which decisions significantly affect fairness, making the evaluation process stronger and more transparent.

The work demostrates the utility of multiverse analysis through a case study on predicting public health care coverage for vulnerable populations. The results reveal that evaluation choices can lead to vastly different fairness outcomes for the same model. This variability is concerning because it allows for the possibility of manipulating metrics to falsely present a biased model as fair.

This multiverse analysis method addresses this issue by highlighting the impact of different design and evaluation decisions on fairness, thereby preventing potential misuse and promoting more transparent and fair ADM systems.



Steps to conduct a multiverse analysis for algorithmic fairness.

TODAY'S STUDENTS, TOMORROW'S FUTURE

relAI trains future generations of AI experts who combine technical brilliance with a keen perception of AI's implications for society. The unique educational concept builds on the observation that a successful education in AI should not be restricted to a single aspect of research, but requires access to an entire spectrum of different qualification measures including scientific knowledge, business expertise, and industrial exposure. Therefore, interdisciplinary training prepares candidates for both academia and industry alike.

The key components of the educational program are:

Personalized Learning: At relAI, education is tailored to individual needs. Each student develops an Individual Development Plan (IDP) which guides their academic journey. The supervision concept ensures continuous mentorship from academic and industry professionals, fostering professional development and focused research.

Early Exposure to Real-Life Challenges: relAI prioritizes practical experience. Students engage in industry internships early in their education, applying theoretical knowledge to real-world problems. This exposure not only enhances learning but also bridges the gap between academia and industry, preparing students for future careers.

International Scope: relAI places a strong emphasis on international exposure. A two-way visiting program allows doctoral researchers to gather experience at top AI centers worldwide, and integrates international fellows into the research and education agenda. This international dimension not only broadens students' perspectives but also enhances their ability to tackle AI challenges on a global scale. The diverse international student community further enriches the learning environment.

»We are equipping the next generation of AI experts with the knowledge and skills they need to become the AI experts of tomorrow.«

Focus: the relAI Curriculum

The school's curriculum is designed to build a comprehensive foundation in AI while addressing real-world applications and ethical considerations. It takes full advantage of the complementary expertise at partnering institutions, and enables a flexible integration into the respective study program.

Its components are:

Lectures & Seminars: Core courses cover mathematical and algorithmic foundations essential for AI research. Specialized modules focus on application domains such as healthcare, robotics, and algorithmic decision-making, ensuring students are well-versed in cutting-edge technologies and trends.

Professional Development Training: Beyond technical skills, relAI emphasizes the importance of professional development. Training in ethics, science communication, and entrepreneurship prepares students to lead in the AI field, ensuring they are not only skilled researchers but also responsible and effective communicators.

Looking back: Event Highlights

Students at relAI not only benefit from our clear structure and extensive network, but also from a number of events throughout the year where they can meet, socialize, and present their work.

WELCOME DAYS

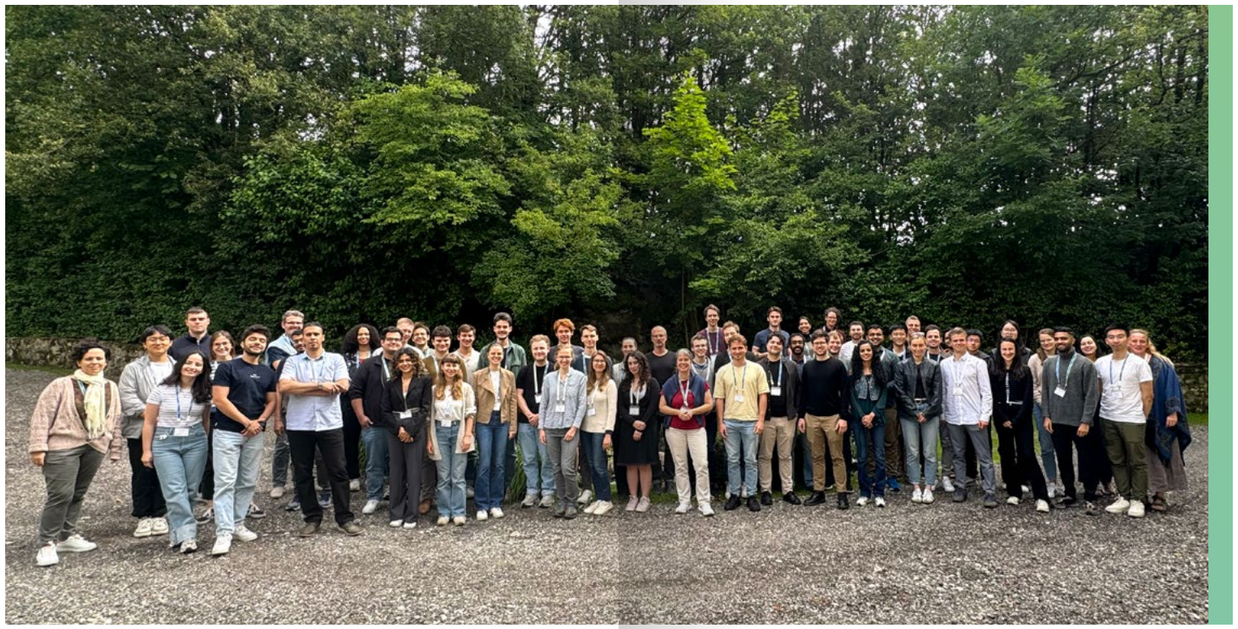
Our Welcome Days at the start of the fall semester provide a warm welcome to the new cohort, introduce them to the relAI world, and connect new and old members in an energizing social gathering.

STUDENT SEMINAR

Our students meet in a biweekly seminar to present their work. The seminar is organized by the student representatives.

MUNICH AI LECTURE SERIES

The Munich AI Lectures are a joint initiative of all major AI initiatives in Munich, coordinated by baiosphere. The lectures take place approximately once a month, and feature insights and ideas from experts.



HACKATHON

In 2024, the first relAI safety hackathon took place. Over the course of a weekend, participants delved into practical projects aimed at addressing various aspects of AI safety and also engaged in thought-provoking discussions surrounding the ethical implications and potential risks associated with AI.

GRAND OPENING

In July 2023, the Konrad Zuse School celebrated its grand opening. Representatives from politics, academia and industry got to know the school, its successful first steps and research conducted by relAI students in a poster session.

RETREAT

A big highlight of the relAI year is the annual retreat, which took place for the third time in 2024. The get-together promotes intense scientific discussion across various aspects of reliable AI and strengthens the community.

STUDENT SOCIALS

Our students regularly meet for various social events, such as a game night or our Christmas party.

RELAI WITH WOMEN IN AI & ROBOTICS

In an engaging community event co-hosted by WAIR (Women in AI & Robotics), our director Gitta Kutyniok talked about Reliable AI: From Mathematical Foundations to Neuromorphic Computing.

ZUSE SCHOOL MEETING

The three Konrad Zuse Schools meet yearly to strengthen their collaboration and discuss upcoming challenges. The Zuse School Meeting took place in Dresden 2023, followed by Munich in October 2024.

AI SAFETY DISCUSSION

In collaboration with AI Safety groups across Germany, we co-hosted Jan Kirchner from OpenAI, who delved into the crucial topic of AI Alignment. After an introductory talk on the topic by a relAI student, Jan Kirchner shared invaluable insights during his presentation and engaged in an insightful Q&A session.

INDUSTRY EVENT

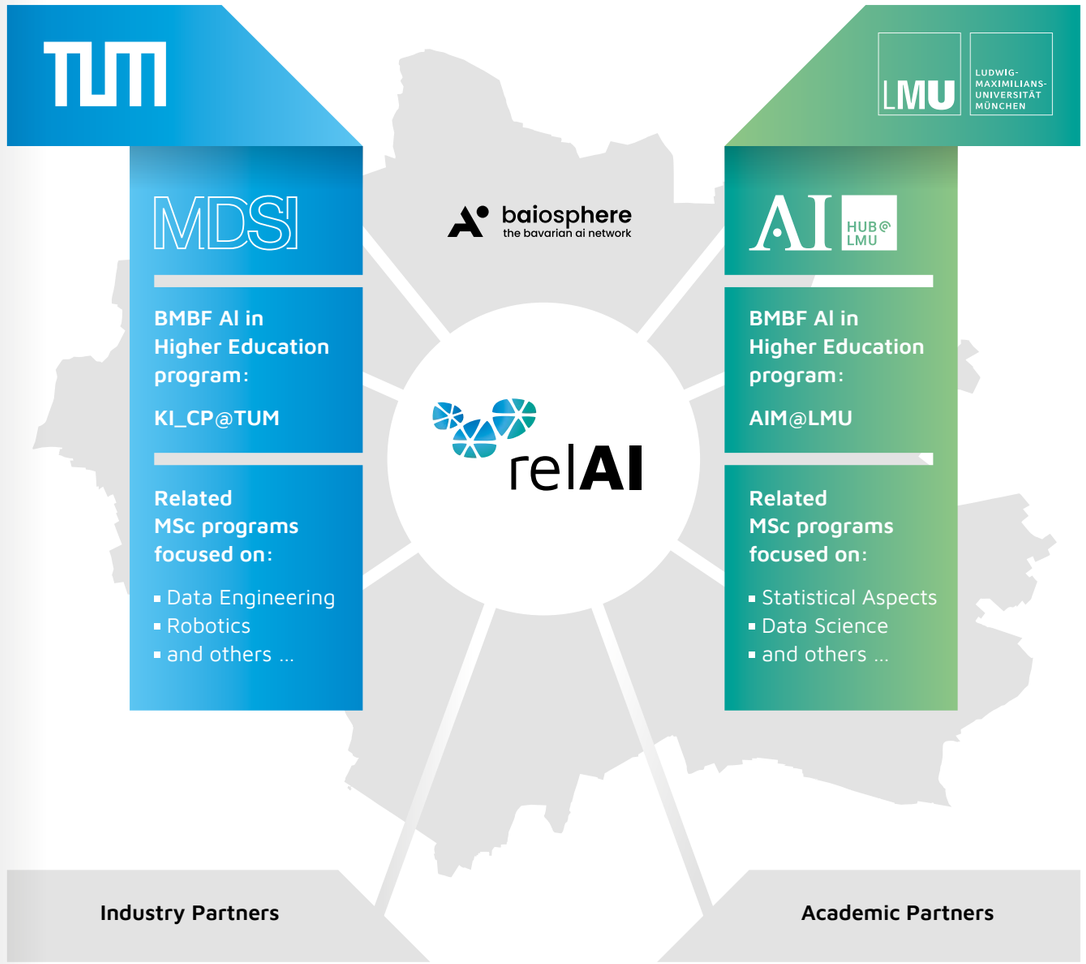
Our relAI students presented their research to the relAI industry partners in a series of industry workshops. Following short lightning talks, intriguing discussions around the reliability of AI took place in smaller breakout groups. Another version of industry events included industry representatives presenting their work to the students.

Building Strong Connections

As outlined before, relAI comprises a network of international academic partners from top international AI centers as well as experts from leading industry partners. In addition to these internal connections, relAI is embedded in the local network in Bavaria and particularly in Munich through its Bavarian AI umbrella organization, baiosphere.

The Munich AI Ecosystem is a vibrant and multi-faceted hub, integrating research, education, and practical applications across various domains. This ecosystem is remarkable due to its diverse components, which collectively foster advancements in artificial intelligence and related fields.

The Konrad Zuse School has become an important part of this ecosystem, forming a link between its various components through its four Research Focus Areas. relAI is bringing together existing



institutions like the Munich Data Science Institute (MDSI), the AI-HUB@LUM, and the Munich Center for Machine Learning (MCML) in the field of Mathematical and Algorithmic Foundations, the LMU Klinikum and the Klinikum Rechts der Isar for Medicine and Healthcare, the Munich Institute for Robotics and Machine Intelligence (MIRMI) in Robotics and Interacting Systems, as well as the Munich Center for Mathematical Philosophy (MCMP) in the field of Algorithmic Decision-Making. relAI is also building bridges in education, linking programs like KI_CP@TUM, AIM@LMU as well as various MSc programs in Data Engineering, Robotic, Statistical Aspects and Data Science.

Engaging with the Public

relAI has undertaken a variety of outreach activities to engage with a broad audience, including current and prospective students, fellows, and the general public. Since launching its website in November 2022, the school has provided comprehensive information about its mission, members, and application processes. The website's global reach is evidenced by a diverse international applicant pool.

The student blog serves as a platform to share cutting-edge research as well as developments at our school, highlighting the significant strides relAI is making toward making AI systems safer, more trustworthy, and privacy-preserving. It is also training the relAI students in their scientific communication skills. The blog is maintained by the students, who organize editorial as well as regular postings by different authors.

In terms of social media presence, relAI is active on LinkedIn and X (formerly twitter), with >40,000 impressions just in 2023. relAI has also gained recognition through various press articles. The press service of both TUM and LMU as well as external media have covered significant events such as the Grand Opening.

WWW X LINKEDIN BLOG

Sustainability Concept

AI and especially the reliability of AI will be an important topic for many years to come. Hence it is our goal to make the Konrad Zuse School an institution that is here to stay.

Quite a number of measures will be used to foster sustainability of the school, and we have already started to implement most of them during the first two years of our existence.

relAI family: We strongly and actively promote a community feeling of the members of our Zuse School in the sense of a "relAI Family." Numerous events as already described in this report bring all members of relAI together on a regular basis, from students to fellows and partners.

Alumni Network: Even though the first students have yet to graduate from relAI, the formation of an alumni network is already in the works. A student group with the support of fellows and the management team is collecting ideas and will establish structures to integrate future alumni into the relAI family.

High Quality Publications: As described in the research section of this report, relAI members have already published a number of excellent papers. Thus, the school can show that its research output is well received in the research community, and the relAI members will continue to aim for this high bar.

Growing Network of Fellows and Partners: relAI receives numerous applications of fellows and industry/academic partners alike who are interested in joining, contributing to, and benefit from our network and education program.

Science Communication: Part of relAI's education program as well as part of making the research visible is the relAI blog, where students write about their work on a regular basis.

Expanding Application Areas: As AI technology is adopted in more and more fields of application, relAI will continue to evaluate and add application areas. In a first step, new relAI Fellows Prof. Kuhn and Prof. Kasneci are exploring AI in Education as an Application Area.

The numbers behind relAI

67

In summer 2024, a total of **67 MSc and Doctoral students** formed the heart of relAI. The school has attracted top international talent and through its structure and community, it supports exceptional young minds in their education, combining technical brilliance with awareness of the importance of AI's reliability.

40 MIO

The Zuse schools were established as part of the AI action plan of the German Federal Ministry of Education and Research ("BMBF-Aktionsplan Künstliche Intelligenz"). **Together, all Zuse Schools receive €40 million in funding.**

2022

relAI was founded in 2022 and welcomed its first students in the same year. It celebrated its grand opening a year later in 2023.

19 companies partner with relAI. Close ties to industry emphasize the transdisciplinary character of the school. The industry partners engage with the students in regular networking events, discussing application-focused challenges. Furthermore, representatives from the industry partners serve as mentors on the transdisciplinary thesis advising committee (TTAC) of the relAI Doctoral students.

19

The members of the Zuse school have published their research in **over 70 publications**, including prestigious venues such as such as NeurIPS, CVPR, and Nature.

70+

The international character of the school is reflected in the fact that **our relAI students represent 21 different countries.**

21

3

There are three Konrad Zuse School of Excellence, all funded by DAAD. Our "sister" Zuse Schools are ELIZA and SECAI. ELIZA is closely connected to the European Laboratory for Learning and Intelligent Systems (ELLIS), a leading European network for AI research. SECAI stands for School of Embedded Composite Artificial Intelligence and is based at the Technical University Dresden and University of Leipzig.

45

45 fellows conduct courses for the students, provide supervision of the doctoral researchers, and jointly conduct research together.

The budget allocated to relAI is €13 million, ramping up from €0.6 million in 2022 to €3 million in 2025.

13 MIO

More than 10 events took place in the academic year 2023–2024. relAI events include talks from world-class experts, community events such as the Welcome Days, our retreat, and even a student-organized hackathon.

10+

40%

The MSc call for 2024 attracted one-third female applicants. Among the selected MSc scholarship holders for 2024, a **total of 40% are female**, highlighting gender diversity.

The social media channels of the graduate school generate **more than 1000 followers** across the platforms LinkedIn and X. Following relAI on social media is a great way to stay updated about the school! The accounts feature event invitations, awards, research updates, and other news about the school.

1.000+

500+

As of July 2024, **more than 500 applicants have applied to relAI**. Of these, around 85% are non-Germans, which showcases the international focus of the school.

Researchers at relAI not only do exceptional work in foundations: They bring theory to life, explore opportunities in the real world, and work on applications that will change every-day life as we know it.

COMMITTED TO OUR RESPONSIBILITY

In Search of a Reliable Intelligent Machine

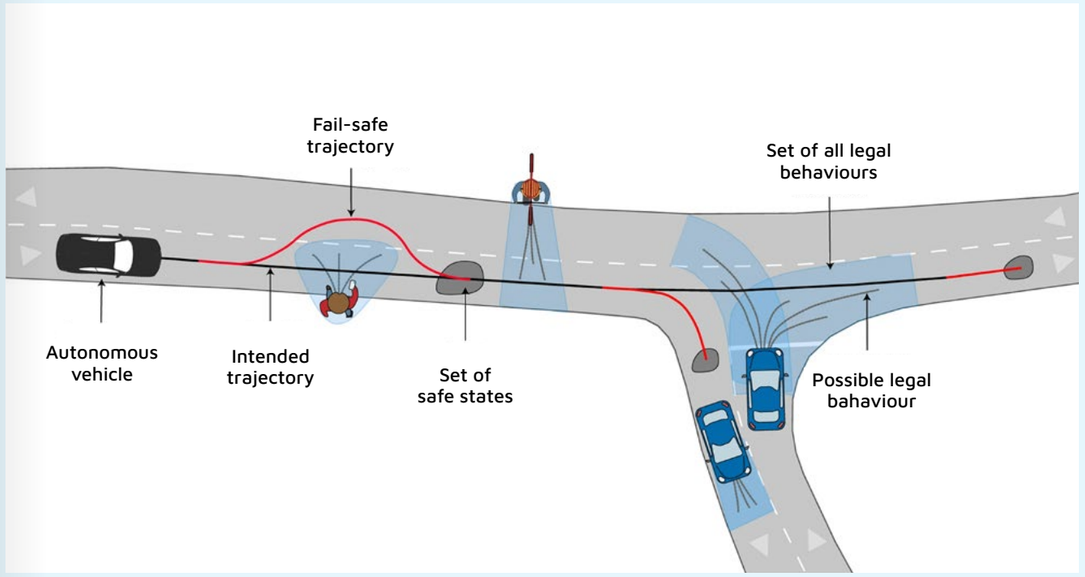
BY DEBARGHYA GHOSHDASTIDAR, relAI FELLOW

The perception of Artificial Intelligence (AI) is often influenced by science fiction, providing narratives on how robots, cyborgs and other intelligent machines would act in their own interest. Among the popular AI characters, C-3PO from Star Wars would be an example of a reliable machine, who acts in the interest of the "good guys," while Agent Smith from The Matrix would perfectly fit the definition of an unreliable AI which plans to destroy both humans and machines. In reality, however, making AI reliable or safe often boils down to preventing AI systems from making stupid decisions that can be unintentionally dangerous. Take the incident that led to the downfall of Cruise robo-taxi.[1] A pedestrian was hit by a car and fell in front of a Cruise driverless car. The car detected an accident and pulled over, but tragically dragged the pedestrian. How did AI fail? Although the AI car detected the accident, it was too simplistic to realize that it was driving over a person. In a different context, researchers found that ChatGPT would spell out plans to destroy humanity or build a bomb if one says "please" in the correct way—if one adds specific characters after the question.[2] To be fair, C-3PO is also bad with keeping secrets.

How do we make AI reliable? A systematic approach is grounded in the framework of adversarial robustness, where the goal is to ensure that the decision of the AI system remains unchanged when the given information changes minimally — for example, how much can a picture of a dog be changed so that AI still identifies it as a "dog"? But, such simplified frameworks are not enough for ensuring reliability of AI in safety-critical systems. Matthias Althoff, Professor for Cyber-Physical Systems at TUM and Fellow at relAI, gives an interesting example: a self-driving car ran over a cyclist crossing the street. Why? Well, because it was only trained using videos of cyclists going in the direction of traffic and pedestrians crossing the street. Within the framework of robustness, one could ask: How much should a video of a cyclist in traffic change to make it look like a cyclist crossing the street? The answer is not straightforward. Prof. Althoff uses formal methods and in particular reachability analysis to certify the safety of autonomous vehicles, which are implemented in the CommonRoad framework developed by his team.[3,4]

Another aspect of reliability is the ability of a system to explain its decisions. Can a driverless car explain how it differentiates a pedestrian from a cyclist? This is becoming more relevant in the era of Large Language Models (LLMs), where people take advice from ChatGPT in daily matters, including therapeutic support.[5] Imagine getting therapy from someone who cannot explain their advice and who could potentially be tweaked to give bad advice like "destroying humanity"! At a recent meeting of the relAI members, there was an engaging discussion on how to make LLMs explainable. Since LLMs are believed to be a step towards Artificial General Intelligence (AGI), building a LLM that is robust and can explain its actions could be the "reliable intelligent machine" we are looking for. While we continue our search, it makes sense to learn from the philosophical nuances of AI reliability that science fiction provides. For instance, Isaac Asimov's robots are often conflicted between the three laws of robotics [6] — the principles of reliability embedded in them.

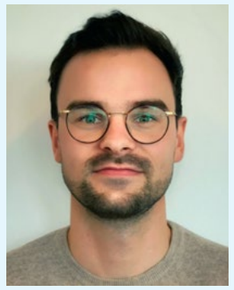
[1] https://fortune.com/2024/05/16/inside-gm-cruise-self-driving-car-accident-san-francisco-what-really-happened/
[2] https://ia.acs.org.au/article/2023/-chatgpt--help-me-make-a-bomb-.html
[3] https://commonroad.in.tum.de/
[4] Pek, C., Manzinger, S., Koschi, M., & Althoff, M. (2020). Using online verification to prevent autonomous vehicles from causing accidents. In Nature Machine Intelligence 2, (pp. 518–528). https://doi.org/10.1038/s42256-020-0225-y
[5] https://chatgpt.com/g/g-aSC9dIv0z-therapist-gpt
[6] https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

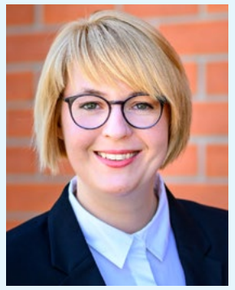


Reachability analysis and verification technique ensure that an autonomous vehicle is safe in accordance with legal safety by maintaining fail-safe trajectories (red lines) at all times. These fail-safe trajectories are collision-free against the set of all legal behaviors (blue areas) of other traffic participants.[4]

Real-World Relevance of Reliable AI

DR. TOBIAS MÜLLER, INDUSTRY-UNIVERSITY COLLABORATION, SAP LABS MUNICH
JONAS KOLK, AI SCIENCE & TECH STRATEGY LEAD I SAP CTO AI OFFICE
DR. KATHARINA WOLLENBERG, INDUSTRY-UNIVERSITY COLLABORATION, SAP LABS MUNICH





We have asked Dr. Jonas Kolk, Dr. Tobias Müller and Dr. Katharina Wollenberg from relAI partner's SAP to give us an insight into use of AI and relevance of reliability in their company. Here is what they told us:

Where does SAP use AI?

SAP uses embedded AI capabilities based on proprietary deep learning and machine learning algorithms to drive business value across SAP's cloud solution portfolio in a relevant, reliable, and responsible way. This allows us, for example, to automatically match invoices to payments in our CloudERP SAP Cash applications, to automatically process invoices in our Spend Management and Business Network product SAP Concur, or to use AI to support content-based integration in the SAP Business Technology Platform. Our AI Foundation is built on the SAP Business Technology Platform, and SAP's AI co-pilot Joule understands business processes across all products. In addition, we foster vibrant AI ecosystem partnerships and investments with leading research institutions and companies.

We also use AI to help users focus on human interactions rather than administrative tasks. For example, AI-powered HR solutions help managers streamline the process of creating annual targets, developing interview questions while removing bias, or creating job postings. The latter can save managers up to two hours per job description, leading to cumulative productivity savings of up to €400,000 for a company with 50,000 employees and thousands of vacancies each year.

Why is AI reliability important to SAP?

SAP specializes in Business AI: AI that is embedded in enterprise business applications and processes from day one. These applications follow SAP's rigorous product standards, AI ethics, privacy, and security standards. SAP customers trust us to run their most business-critical processes, ensuring the reliability, accuracy, precision, and availability of SAP Business AI. Potential failures can negatively impact businesses and end users. Since we recognize the profound impact of AI on decision-making, fairness, transparency, and privacy, we address issues of bias and discrimination in the development of our AI-based applications and work to ensure transparency and explainability of the AI-based system's actions.

How does SAP collaborate with universities, particularly in the area of incubating startups?

SAP is collaborating with research institutes around the world to explore the application of AI technologies in product and business process innovation, and the impact of AI on the world of work.

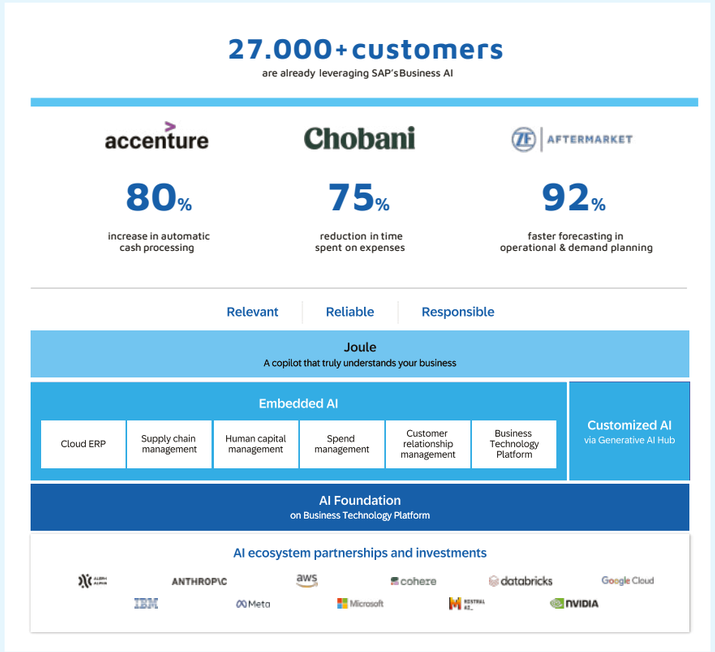
In collaboration with UnternehmerTUM, for example, SAP has gained extensive experience in founding and supporting AI start-ups. UnternehmerTUM offers start-ups a full range of services, from the initial idea to the IPO, and supports them with a team of more than 300 employees, as well as venture capital financing. Particularly noteworthy is their engagement in the field of AI through formats such as XPLORE, XPRENEURS or TECH Founders, where interdisciplinary start-up teams can develop their product within a few months, bring it to market maturity, or start the scaling process. In this context, the Technical University of Munich plays a central role in the ecosystem by contributing talent and research results which are further developed and commercialized through the Unternehmer TUM platform. This collaboration advances the development of AI technologies and strengthens the innovation network of SAP and its partners. Given our focus on relevant, reliable, and responsible AI, we highly value these partnerships with various (external) experts, such as relAI or UnternehmerTUM with their network of experts, students, and professors.

How does SAP benefit from being a relAI partner?

At SAP, we are committed to helping our customers leverage AI to create tremendous business value. We specialize in Business AI: AI that is relevant since it is embedded in enterprise business applications and processes from day one; AI that is reliable because we train, ground, and adapt AI on companies' business data and context; and finally, AI that is responsible by design because it follows SAP's rigorous ethics, privacy, and security practices.

With regard to these aspects, partnering with relAI provides SAP with strategic and operational benefits. By collaborating with relAI, we can engage with cutting-edge research in the areas of AI reliability, safety, security, and privacy-preservation. This opens up opportunities for joint research projects, enabling the development of proprietary innovations and solutions. Moreover, relAI provides a pipeline of highly skilled AI experts trained to uphold the highest standards of AI reliability, giving us access to the next generation of AI leaders.

Overall, the partnership not only consolidates our portfolio of AI-related applied research projects but also fosters a more profound knowledge exchange and talent engagement around the development of reliable AI systems and other Business AI topics.



Inside relAI Student Perspectives

LISA WIMMER,
DOCTORAL STUDENT
FROM GERMANY

How does relAI help you in building networks?

A vibrant network is valuable in any job but indispensable in scientific research. For one, exchanging ideas with like-minded people helps identify research gaps and initiate collaboration on promising ideas. relAI connects different fields across all levels of experience and organizes meeting opportunities where the relevant people come together. For me, this has resulted in three joint articles with relAI affiliates in the past year. Further down the road, having personal contacts in both academia and industry will be immensely helpful for shaping my further career.

How does relAI support you in disseminating my research?

Even the best work will remain without impact if it's not broadcast into the community. For me as a young researcher, it's all the more challenging to disseminate my work among so many peers in the field. In presentations, poster fairs and lightning talk sessions organized by relAI, I've had the opportunity to speak about my research and, most importantly, discuss it with the right people.

How do you experience relAI as a community?

relAI is an incredible and closely-knit academic community centered around a shared passion for AI. It's one of the largest graduate schools in Bavaria, attracting a diverse mix of students and faculty. The program includes both MSc and Doctoral students, which really creates a multidisciplinary environment where we can learn from different perspectives within AI. The involvement of well-known professors from various fields is a huge bonus, as it gives us early access to a wealth of knowledge and expertise within the community. What I really appreciate about relAI is how connected the community feels. There are always activities going on, like industry talks and workshops. It's also great to see students taking the initiative, e.g. to organize our own seminars and hackathon, which adds to the dynamic and collaborative vibe of the place. Based on my personal experience working on industry excursions and invited talks, I have seen how we get support from the fellows and directors, which helps us build valuable connections. And then there's the yearly retreat, which offers a chance for intensive discussions and helps deepen the bonds within the community. All in all, relAI provides a well-rounded experience, combining rigorous academic training, practical industry insights, and a close-knit community.

How does relAI influence the focus of your degree?

relAI has had a significant influence on how I've geared my degree to my interests. Thanks to the financial support relAI provides, I've been able to concentrate more on lab work and immerse myself in research from an early stage. This hands-on experience has been invaluable in shaping my understanding and interests within the field. Additionally, the mentoring and supervision I receive through the Individual Development Plan have been crucial. Mentoring has allowed me to connect with Doctoral students and professors, giving me deeper insights and helping me think more critically about my academic and career goals. I'm currently studying Computational Linguistics, and relAI's influence has broadened my perspective, particularly regarding the importance of AI/LLM safety and alignment issues. This is an area I hadn't considered much before, but now I'm keen to explore it further in the latter half of my MSc program. The guidance and resources provided by relAI have not only enhanced my current studies but also helped me identify new and exciting areas that I want to pursue.

CHENGZHI (MARTIN) HU,
MSC STUDENT FROM CHINA

»relAI connects different fields across all levels of experience and organizes meeting opportunities where the relevant people come together.«

»I advise all new MSc students at relAI to actively participate in these events.«

AMINE KETATA,
MSC STUDENT FROM TUNISIA

How can you get the most out of relAI as an MSc student?

There are plenty of opportunities for MSc students at relAI for self-development as well as to learn more about the work of other relAI members. At relAI, we organize weekly seminars where students regularly share their recent research or any other work in the field of reliable AI that they find interesting. In addition, relAI offers us the opportunity to take part in courses at both TUM and LMU. We also have a yearly retreat, where students and fellows spend three days together, which is an incredible chance to connect with potential collaborators or mentors. I advise all new MSc students at relAI to actively participate in these events.

What career options do you have once you graduate from relAI?

After graduating from relAI as an MSc student, you have the opportunity to pursue a Doctoral within relAI at one of the affiliated chairs. This is a natural transition and allows students to continue working on the research they started, for example, during their MSc theses. On the other hand, relAI graduates have a great chance of landing very good jobs, as reliable AI methods are increasingly sought after. The industry network of relAI serves as a great network for a job search.

Get involved

(Prospective) students:
We usually publish our call for MSc applications in the spring and our call for Doctoral students with excellent transcripts in late fall. You can join our mailing list to get informed about the start of the application period.
We expect a bachelor's (and MSc) degree (or equivalent) in computer science, mathematics, engineering, natural sciences or other data science/machine learning/AI related disciplines, a high grade point average, and a genuine interest to work on a topic of reliable AI.

Industry partners:
We are always looking for industry partners working on reliable AI. Please reach out at coordinators@zuseschoolrelai.de to discuss a collaboration.

Politics, Press and General Public:
You can contact us via our general mail: coordinators@zuseschoolrelai.de.

LEGAL NOTICE

Responsible party in terms of press law:
relAI Directors

Prof. Stephan Günnemann
Munich Data Science Institute (MDSI)
Technische Universität München
Walther-von-Dyck-Straße 10
85748 Garching, Germany

Prof. Dr. Gitta Kutyniok
Ludwig-Maximilians-University München
Akademiestraße 7
80799 Munich, Germany

Editorial Team
Dr. Mónica Campillos, Daria Kozlova,
Maria Matveev, Nora Ott,
Dr. Andrea Schafferhans

Graphic design
grafikcafé :: feines design

Photo credits
The photos were provided with the kind permission of:
StMWK/Pablo Castagnola,
TUM/Astrid Eckert, TUM/Andreas Heddergott,
DAAD/Paul Kuchel Pykado, Conny Mirbach,
Ansgar Pudenz and TinoBoecher

relAI
Konrad Zuse School
of Excellence in Reliable AI

Munich Data Science Institute (MDSI)
Walther-von-Dyck-Str. 10
85748 Garching, Germany

info@zuseschoolrelai.de
www.zuseschoolrelai.de

Technical University of Munich

LMU Ludwig-Maximilians-Universität München

SPONSORED BY THE
Federal Ministry of Education and Research

DAAD Zuse Schools
Konrad Zuse Schools of Excellence in Artificial Intelligence