1 Introduction

The current technological revolution is largely driven by spectacular progress in artificial intelligence (AI). Yet, although the huge potential is widely recognized, the lack of **reliability of AI technology** is still considered a serious issue of concern, limiting its adoption both by industry and society at large. Indeed, aspects such as **safety, security, and privacy-preservation** are essential prerequisites for the use of AI in domains of public interest[1] including application areas such as autonomous driving and human-robot interaction, healthcare, or decision-making for business optimization. Since these are all crucial innovative fields of German key industries, the need for reliable AI technology has also been identified by the German Federal Government with its AI Strategy aiming to promote research regarding reliability, explainability, and accountability of AI[2].

While the research community has recently started to address the lack of reliable AI technology, it still remains a largely underdeveloped area of teaching and research. In particular, we are missing trained talents in the field of reliable AI having obtained the skills to transfer research results in this field to corresponding applications in industry. Accordingly, many companies in Europe and Germany, as well as many public organizations lack the technical knowledge to develop reliable AI systems. A proper integration of reliable AI into the higher education system is, thus, highly required to ensure that Germany catches up and fills the role as a leading nation in the growing market of AI.

The vision of the proposed “Konrad Zuse School of Excellence in Reliable AI” (relAI) is to train future generations of AI experts, who for the first time **combine technical brilliance with awareness of AI’s implications on society**. Our novel, highly innovative AI program will educate top international candidates in the **end-to-end development of reliable AI systems** - covering the full spectrum of scientific knowledge, business expertise, and industrial exposure. Our program will prepare students and doctoral candidates for diverse career paths in industry and academia. Highlights of our training program include collaborations with and work experience in top international AI centers combined with a strong interaction with industry partners to foster **high impact interdisciplinary** research**.** We expect reLAI to become a lighthouse initiative in AI which will attract top talents from abroad and thus strengthen Germany as a leading country in AI.

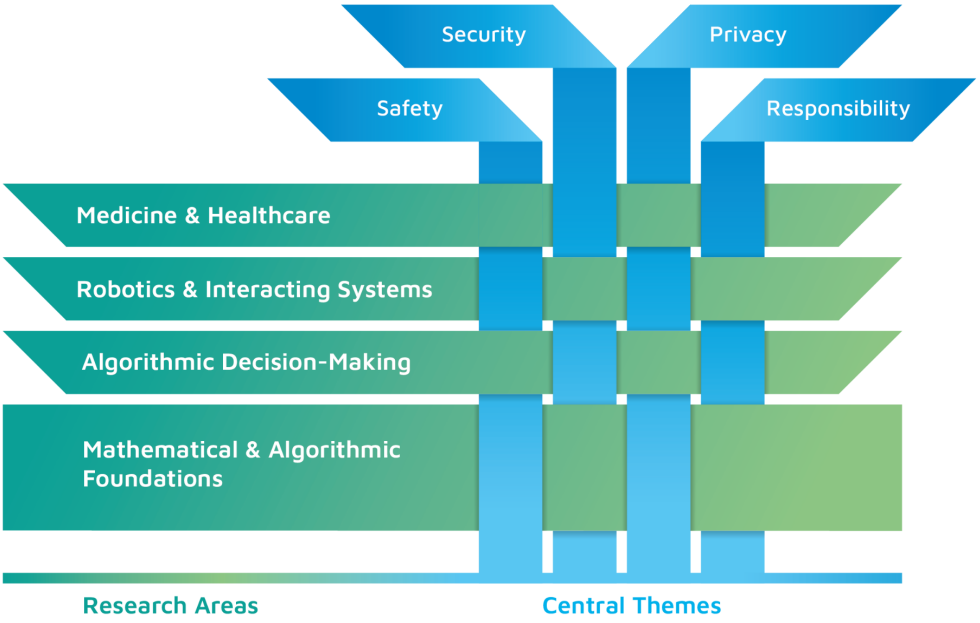
The school will be embedded into the unique transdisciplinary Munich AI ecosystem, a hotspot in Germany for AI development, where it combines the expertise of the **two Universities of Excellence** TUM and LMU – which host numerous well-known academics in areas of central importance to this proposal including foundational AI, robotics and interacting systems, medicine and healthcare, and decision-making – with multiple leading **industry partners** and **non-university research institutions.** Specifically, two Fraunhofer Institutes (AISEC, IKS), focusing on secure and safe AI, respectively, act as a bridge to transfer the obtained research to respective applications. Combined with the continuing support of Bavaria's High-Tech Agenda, with multiple further AI professorships to be established in the upcoming years, the School will flourish on fertile ground.

2 Research Focus

As its central theme, the scientific program of relAI will contribute 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 relAIl: Our conception of reliability involves the demand for a rigorous formal description of properties as well as provable guarantees, because only such guarantees will create the trust and confidence needed for an unreserved adoption of AI in practice.

The thematic structure of relAI is visualized in the figure below: The research program combines mathematical and algorithmic foundations of reliable AI along with domain knowledge in three core **application domains**: 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 domains of public interest.

[1] Winter et al. Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications, 2021
[2] Artificial Intelligence Strategy of the German Federal Government, the Federal Government, 2020



In this regard, each of the school's four research areas (green in figure above) covers central themes of reliable AI (blue in figure above):

- **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 against adversaries.
- **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 taking societal norms, ethical principles, and the need of people into consideration, for example by making decisions understandable and protecting individuals against discrimination.

The structure of relAI enables every student and researcher to cover the entire spectrum of reliable AI topics while still being able to focus on specific domains of interest. Moreover, by combining foundational AI research with core applications, we realize a strong interdisciplinarity within relAI, which in turn allows us to build a tight link to the various industry fellows contributing to these areas and ensuring the impact of our research in real-life settings. Overall, the students and researchers will be exposed to a breadth of topics and offered attractive qualification possibilities and job perspectives in multiple fields of AI. In the following, each area is described in more detail.

2.1 Mathematical & Algorithmic Foundations

Reliability of AI with all its facets can only be achieved through a profound understanding of its foundations. In fact, the current 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 aim to address are twofold. Firstly, we aim to establish **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 are novel and highly complex architectures such as GraphNNs 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.

Tackling these challenges requires expertise in multiple fields, which is represented, among others, by our fellows Althoff, Bischl, Böhm, Drton, Eckert, Ghoshdastidar, Günnemann, Hirche, Hüllermeier, Kauermann, Kilbertus, Kutyniok, and Tresp. Examples of research topics include:

- **Safety:** We investigate how to incorporate hard constraints on the AI system's behavior and the awareness of its own uncertainty to ensure safe use[3]. Adequate theoretical foundations on the basis of formal logic and probability theory (and their combination) should lead to certification/verification with provable guarantees[4]. Exemplary thesis topic: *Efficient Robustness Certificates for Transformer Architectures*.
- **Security:** Modern AI and machine learning (ML) add further vulnerabilities to computing infrastructure, for example due to the reliance on training data that can be manipulated or the black-box character of many ML methods[5]. To obey the highest security standards, also for the use of ML in distributed or internet of things scenarios, theoretical foundations in cryptography, security protocols, access control models, etc. need to be extended correspondingly. Exemplary thesis topic: *Reliable lightweight security for resource-constrained machine learning*.
- **Privacy:** The reliance on data and the use of specific computing infrastructure, such as distributed systems based on federated learning, make AI systems also challenging from a privacy point of view. Thus, in addition to the state-of-the-art approach of differential privacy, we study new theoretical foundations to guarantee that privacy is preserved for such systems[6]. Exemplary thesis topic: *Privacy-preserving deep learning*.
- **Responsibility:** Fairness and explainability are of utmost importance in decision-making, such as algorithmic hiring, though their theoretical foundations are still under way[7]. While explainability is even lacking a proper formalization, existing methods on fairness, e.g., based on causal modeling and counterfactual reasoning, are missing provable guarantees and a backing from the social sciences. Exemplary thesis topic: *Subgroup approaches for explainable AI*.

2.2 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[8]. The emergence of AI for medicine and healthcare also offers a number of transformative opportunities for economic growth. Examples cover 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 particular, advances are required in robust and data efficient learning, privacy preservation, and interpretable deep learning.

In relAI, we combine the expertise in AI healthcare and medicine to successfully tackle these challenges, including fellows Bhatotia, Böhm, Buyx, Feuerriegel, Haddadin, Ingrisch, Kaissis, Rückert, Schmidt, Schnabel, Theis, and Tresp. Research topics we aim to investigate include:

- **Safe and Robust Medical AI:** To guarantee the level of safety required by applications in healthcare and medicine, we investigate robust ML techniques that can deal with missing or sparse data as well as variations across datasets. In the context of medical imaging, for example, AI approaches need to be robust to variations in the imaging data (e.g. different scanners) and variations across patients (e.g. different patient populations or pathologies). Exemplary thesis topic: *Learning from sparse medical annotations*.
- **Privacy of Healthcare Data:** In order to harness the large amounts of clinical data that are regularly collected across the healthcare system, it is crucial to appropriately address the challenges of privacy and data protection. Techniques such as federated learning and secure

[3] An, Liu, Zhang, Chen, Chen, Sun. Uncertainty modeling and runtime verification for autonomous vehicles driving control: A machine learning-based approach. Journal of Systems and Software, 167, 2020.

[4] Bojchevski, Klicpera, Günnemann. Efficient Robustness Certificates for Discrete Data. International Conference on Machine Learning, 2020.

[5] Xue, Yuan, Wu, Zhang, Liu. Machine Learning Security: Threats, Countermeasures, and Evaluations. IEEE Access, 8:74720-74742, 2020.

[6] Liu, Ding, Shaham, Rahayu, Farokhi, Lin. When Machine Learning Meets Privacy: A Survey and Outlook. ACM Computing Surveys, 54(2):1-36, 2022.

[7] Kolek, Nguyen, Levie, Bruna, Kutyniok. A Rate-Distortion Framework for Explaining Black-box Model Decisions. In: xxAI - Beyond explainable Artificial Intelligence, to appear (arXiv:2110.08252), 2022.

[8] Topol. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 2019.

multiparty computation offer the potential of learning without moving clinical data out of the hospital or healthcare system[9]. Combined with differential privacy techniques, this can enable privacy-preserving ML on large-scale, distributed patient databases. Exemplary thesis topic: *Differentially private deep learning for medical imaging.*

- **Explainable AI in Medicine**: While recent AI approaches such as deep learning have shown great promise, their foundational principles and internal knowledge representations are still poorly understood, making them difficult to apply in medicine, where it is important to understand cause and effect. We aim to extend principles for explainability in order to ensure that AI solutions in healthcare are trusted by clinicians, patients, and regulators. Exemplary thesis topic: *Learning interpretable treatment policies for medicine*.

2.3 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 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 situations the agent will face in the future? How to balance the advantages of AI cloud computing with the increased risk of security violations? How to leverage data to adapt to the needs of a human user while bearing privacy concerns in mind?

To answer such questions, relAI will focus on safe, secure, and privacy-preserving AI in the context of autonomous agents and interacting systems. Our School represents core expertise on these topics via the fellows Althoff, Bhatotia, Butz, Haddadin, Hirche, Kreuter, Kutyniok, Schmidt, Schütze, and Zöller, among others. Specific research topics we aim to address include:

- **Safety and Certification of AI-enabled robots:** Nowadays, robotic systems are still mostly certified using standardized tests[10]. This procedure, however, is reaching its limits due to the rich decision-making capabilities of AI-enabled robots. Combining reinforcement learning with formal verification allows the certification of intelligent agents for training and deployment[11]. We aim to safeguard reinforcement learning by making use of reachability analysis to guarantee the safety of planned trajectories, specifically also tackling online scenarios. Exemplary thesis topic: *Provably safe reinforcement learning*.
- **Security:** Today, most data-driven applications are supported by cloud providers. While providing unique competitive advantages, the transition to the cloud comes with an increased risk of security violations. In untrusted environments, an attacker can compromise the security properties of the stored data and query operations. To address these security threats, we will investigate hardware-assisted confidential computing, where we provide strong security properties with low trusted computing[12]. Exemplary thesis topic: *Reliable and secure intelligent applications for cyber-physical systems*.
- **Privacy by Design**: While intelligent systems rely heavily on data related to the situation and interaction to adapt to user needs, the exploitation of data is hindered by many constraints, ranging from legal and ethical concerns to brand experience and general reservation towards data-gathering systems. Finding a balance between data usage and privacy is critical for the development of usable and acceptable systems. We will investigate how privacy by design[13], where already in early phases of the system design privacy is a major concern, can be realized in complex interacting systems. Exemplary thesis topic: *Privacy by Design: a human centered experimental investigation of trade-offs in the design of intelligent systems*.

[9] Kaissis et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nature Machine Intelligence 3 (6), 2021.

[10] Haddadin, Albu-Schäffer, and Hirzinger. Requirements for Safe Robots: Measurements, Analysis and New Insights. The International Journal of Robotics Research, 28, 1507–27, 2009.

[11] Krasowski, Wang, Althoff. Safe Reinforcement Learning for Autonomous Lane Changing Using Set-Based Prediction. IEEE International Conference on Intelligent Transportation Systems, 2020

[12] Bailleu, Giantsidi, Gavrielatos, Le Quoc, Nagarajan, Bhatotia. Avocado: A Secure In-Memory Distributed Storage System. USENIX Annual Technical Conference, 65-79, 2021.

[13] Spiekermann. The challenges of privacy by design. Communications of the ACM, 55(7), 38-40, 2012.

2.4 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? Decisions of that kind are increasingly automated and made by algorithms instead of humans, often relying on AI methods. Our ambition is to develop AI-based methodologies for reliable algorithmic decision-making (ADM). This comes with the need to address specific technical issues such as the lack of an objective "ground truth" underlying every prediction, and learning from partial training information, comprising feedback about the decision made, while lacking information about counterfactuals[14]. Methodological research on ADM will be complemented by more application-oriented research on reliable decisions in business and management.

Our school covers these topics with complementary expertise by the fellows Bischl, Brosius, Butz, Drton, Feuerriegel, Günnemann, Hüllermeier, Ingrisch, Kilbertus, Kreuter, List, Schnabel, and Zöller amongst others. Exemplary research directions which we aim to address in relAI include:

- **Privacy**: As AI methods for ADM need access to people's private data, possibly collecting and combining such data from different sources, the privacy of people must be protected. To successfully apply formal methods such as differential privacy or federated learning in the social context, we investigate adaptations and new techniques that account for specific characteristics of data in these settings (e.g., missing data, imputed values, correlated response variance, etc.). Exemplary thesis topic: *Preserving privacy in interactions with super assistants*.
- **Safety and trust:** Decisions impacting society and people must be trustable and safe in the sense of avoiding – as much as possible – serious misjudgments. This implies, for example, that a learning system must be sufficiently aware of its own (in)competence and (un)certainty about the right course of action. This awareness is challenged in many situations, for instance, when models are trained in one context and used in a different one, or if changes happen over time. Exemplary thesis topic: *Epistemic uncertainty in algorithmic decision-making*.
- **Algorithmic fairness**: Methods such as multicalibration[15] have been shown to protect subpopulations from miscalibrated predictions. Yet, the impact of fairness for AI on ADM is little understood. In fact, all evidence is based on numerical experiments, while actual field evidence is rare. Collecting such field evidence will allow the study of various human-centered metrics, such as trust, adherence, and performance. Adapting and testing these methods in applied settings will extend the benefits of evidence-based decision-making to communities that do not have the resources to collect high-quality data on their own. Exemplary thesis topic: *Development of data-driven frameworks for identifying unfairness in algorithmic decision-making.*

2.5 Innovation Potential & Fellows

The thematic focus and structure of relAI forms a unique program within the German scientific and industrial landscape, thereby acting as a lighthouse initiative for Germany worldwide. As reliability is crucial for the adoption of AI in industry – a perception fully confirmed by our industry partners – as well as the development of the scientific field itself, our program has the potential to innovate the progress in various application domains. We would even say that relAI will enable the use of AI in applications inaccessible so far, due to strict regulations or safety concerns. By educating talents across different areas, not only interdisciplinary training but also attractive job prospects will be realized. The success of relAI is largely rooted in a unique composition of fellows, which allows for combining fundamental AI research with strong AI application knowledge. In accord with our goal to foster collaboration between the different partners, this research will be put into practice with the help of our industry partners, leading to high-impact applications and technological innovation (see also Section 3.2). Overall, the "Konrad Zuse School of Excellence in Reliable AI" will thus significantly strengthen the research output of Germany in a highly impactful field of AI.

[14] Hüllermeier. Prescriptive Machine Learning for Automated Decision Making: Challenges and Opportunities. arXiv:2112.08268, 2021.

[15] Hebert-Johnson, Kim, Reingold, Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. International Conference on Machine Learning, 2018.