

# CONSTRUCTING EXPLAINABILITY

## TRR 318

### Imprint / Legal Notice

**Published by:**

TRR 318 Constructing Explainability  
Transregio of Paderborn University and Bielefeld University  
Zukunftsmeile 2  
33102 Paderborn

**Edited by:**

Linda Thomßen (Project Ö, TRR 318)

**Translated by:**

Partially machine-translated and edited by Carly Ottenbreit

**Designed by:**

Sven Carlmeyer (Project Ö, TRR 318)

**Printed by:**

Bonifatius GmbH

**Printing run:**

500 Copies

**Online at:**

[www.trr318.de/en](http://www.trr318.de/en)

At TRR 318, our research is driven by the imperative to make the explanations and decisions of artificial intelligence systems understandable. Over the past four years, the TRR team has achieved several significant milestones, including the development of innovative methods for evaluating the effectiveness of explanations across various applications, the creation of systems that can explain AI decisions, and the incorporation of social perspectives on explainability.

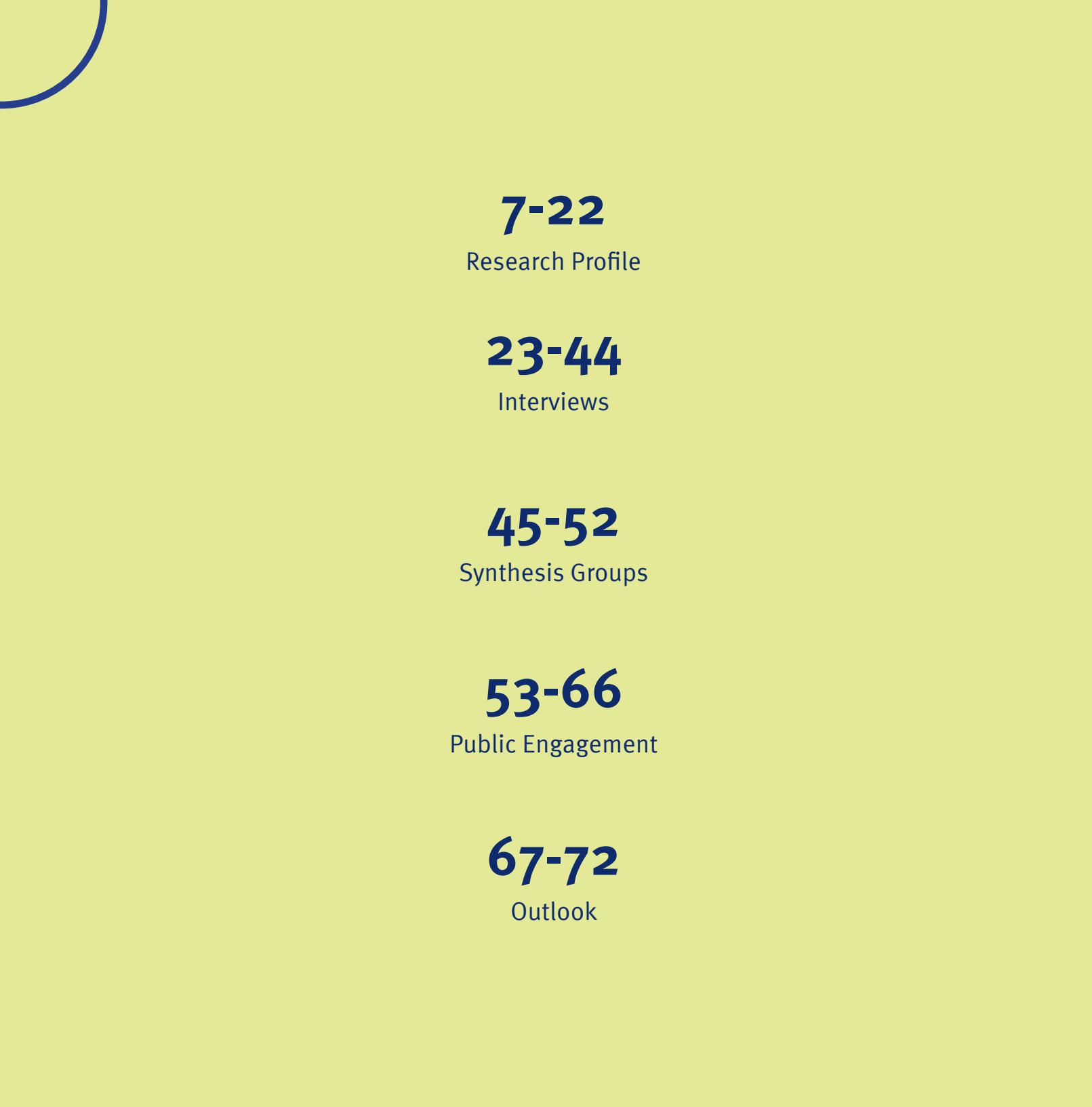
Our view is that explanations are co-constructed, meaning that users who receive an explanation actively participate in the process of explanation. The TRR research aims to make AI more transparent and accessible, encompassing a wide range of interdisciplinary approaches. Effective communication and sharing knowledge among all TRR members are essential in fostering a productive research environment.

In this brochure, we present our research profile and detail a number of our cutting-edge research projects, while also featuring the voices of our TRR members, who share their experiences, goals, and findings from the first TRR 318 funding period.

We would like to thank all the members involved, as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Paderborn University, and Bielefeld University for making this research possible.



Prof. Dr.-Ing. Britta Wrede and Linda Thomßen (in summer 2025)  
Project Ö/ TRR 318



**7-22**

Research Profile

**23-44**

Interviews

**45-52**

Synthesis Groups

**53-66**

Public Engagement

**67-72**

Outlook

STENT  
CON



# RESEARCH PROFILE



Algorithm-based approaches, such as machine learning, are becoming increasingly complex. This complexity – along with the lack of transparency behind these technologies – is making it more difficult for human users to understand and critically reflect on the decisions offered by artificial intelligence (AI). In response to these societal challenges, computer scientists have begun to develop self-explanatory algorithms that provide intelligent explanations, a process known as explainable artificial intelligence (XAI). XAI programs, however, only interact with humans to a limited extent and do not take into account the context or the types of information the human user being addressed needs. This runs the risk of generating AI explanations that will not be understood by the human user.

Members of the Transregional Collaborative Research Center Constructing Explainability (TRR 318) are challenging this view by instead conceiving of explanation as a process of co-construction: in this model, the person receiving the explanation takes an active role in the AI explanatory process and co-constructs both the goal and the process of explanation. Pursuing a collaborative, interdisciplinary approach, the mechanisms of explainability and explanations are being investigated by 22 project leaders, who are assisted by approximately 40 researchers from fields ranging from computer science, economics, and linguistics, to media science, philosophy, psychology, and sociology.

The findings of the research efforts from TRR 318 contribute to the development of:

- Multi-disciplinary understanding of the explanatory process linked to the process of understanding and the contextual factors influencing it.
- Computer models and complex artificial intelligence systems that can generate situation-specific and efficient explanations for their addressees.
- A theory of explanation as social practice that considers the expectations of the person being communicated with, as well as their role in the interaction.

These foundations for explainable and understandable artificial intelligence systems will enable greater active and critical participation in the digital world.



## **Publications**

<https://trr318.uni-paderborn.de/en/publications>

**Research Area A** (explaining process): focusing on the interactive process of multimodally co-constructing an understanding of what is being explained.

**Research Area B** (explanation as social practice): extending the microlevel of explanation to the meso- and macrolevels, i.e. the societal dimensions.

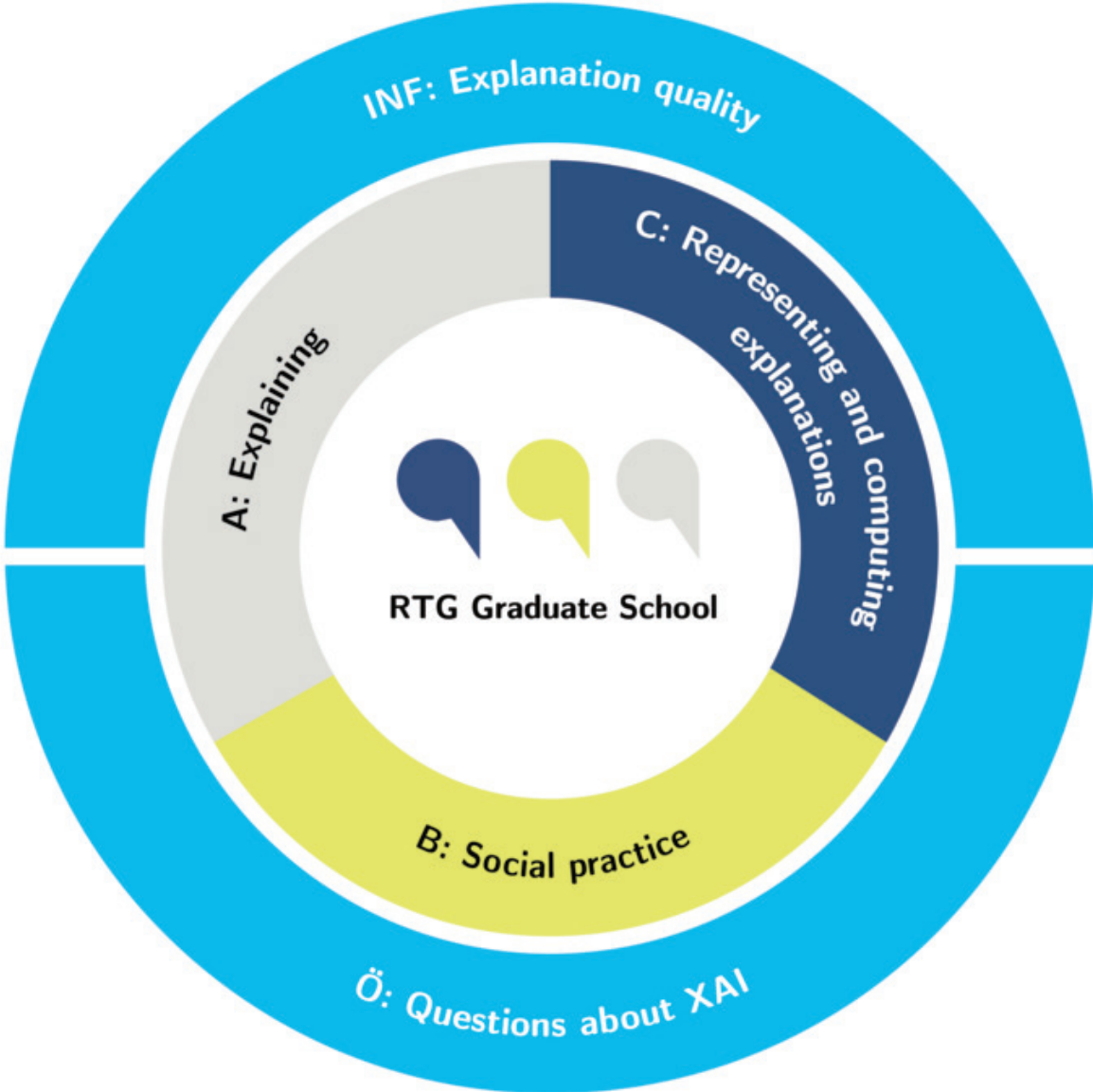
**Research Area C** (explanatory representation): addressing the dynamic properties of conceptualizing the explanandum.

**Project INF** provides an overarching research infrastructure; **project Ö** facilitates public relations and outreach; **project RTG** provides a framework for educating doctoral and postdoctoral researcher; and **project Z** addresses administrative, organizational, and financial matters.

## New Projects and Independent Research Groups

In the first funding period, two new professors, Prof. Dr. Hanna Drimalla and Prof. Dr. Suzana Alpsancar, joined TRR 318 with two novel projects: Ao6 and Bo6, respectively. These projects extend the existing TRR structure along two important directions: Ao6 considers individual differences in the understanding process and Bo6 ethically reflects on the technological developments within TRR 318.

The two new independent research groups, led by Dr. Christian Schulz and Dr. David Johnson, are directly linked to the TRR focus of the co-construction of explanations and complement the work of the Transregio. They are funded by Paderborn University and Bielefeld University, with the aim of supporting outstanding young scientists early on in their careers.



## Area A - Explaining

- Ao1 Adaptive explanation generation
- Ao2 Monitoring the understanding of explanations
- Ao3 Co-constructing explanations with emotional alignment between AI-explainer and human explainee
- Ao4 Integrating the technical model into the partner model in explanations of digital artifacts
- Ao5 Contextualized and online parametrization of attention in human-robot explanatory dialog
- Ao6 Co-constructing social signs of understanding to adapt monitoring to diversity

## Area B - Social Practice

- Bo1 A dialog-based approach to explaining machine learning models
- Bo3 Exploring users, roles, and explanations in real-world contexts
- Bo5 Co-constructing explainability with an interactively learning robot
- Bo6 Ethics and normativity of explainable artificial intelligence

## Area C - Representing and Computing Explanations

- Co1 Healthy distrust in explanations
- Co2 Interactive learning of explainable, situation-adapted decision models
- Co3 Interpretable machine learning: explaining change
- Co4 Metaphors as an explanation tool
- Co5 Creating explanations in collaborative human-machine knowledge exploration
- Co6 Technically enabled explanation of speaker traits

## Intersecting Projects

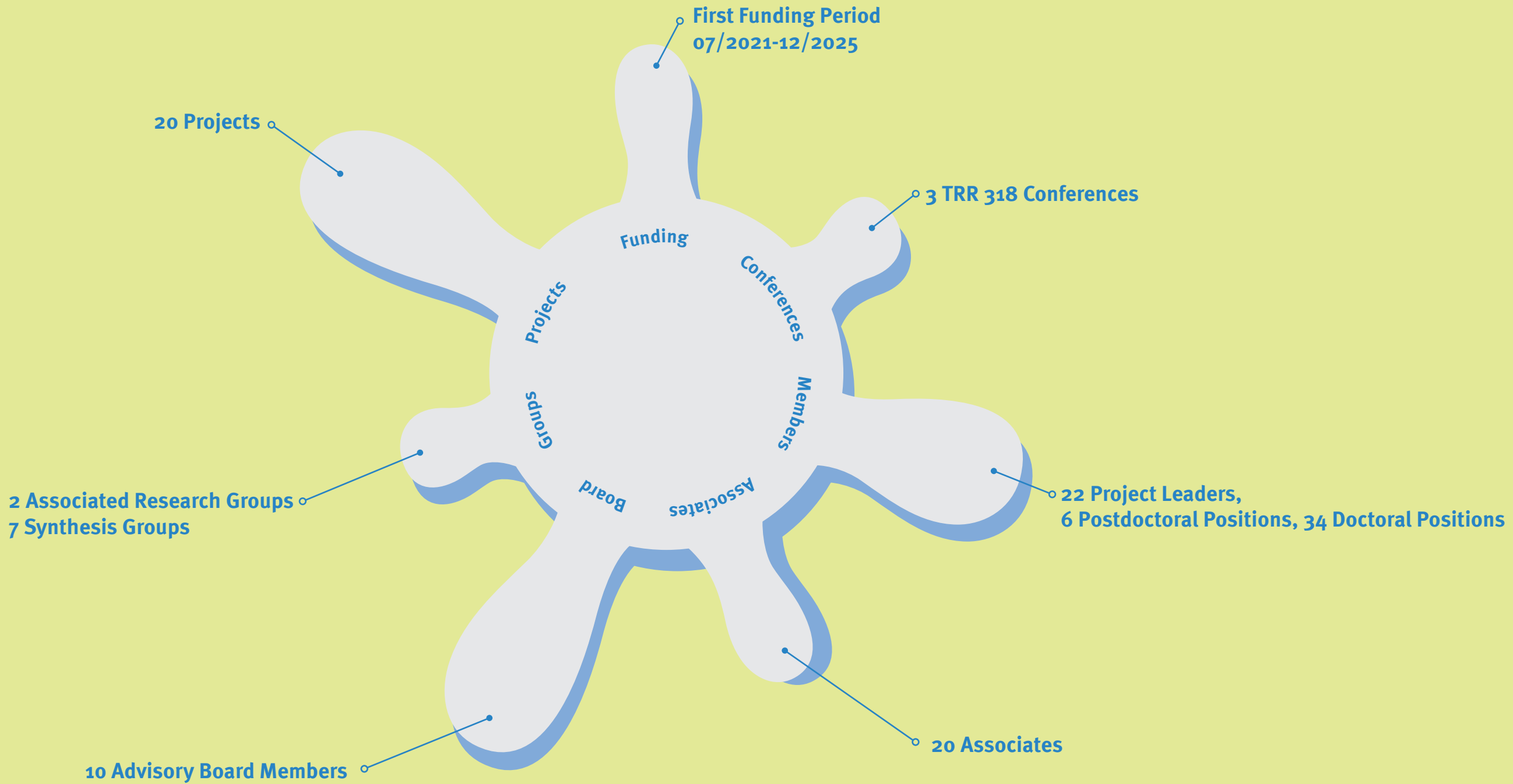
- INF Toward a framework for assessing explanation quality
- Ö Questions about explainable technology
- RTG Research training group
- Z Central administrative project

## Project Leaders in the First Funding Period

Suzana Alpsancar	Heike Buhl	Hendrik Buschmeier
Philipp Cimiano	Hanna Drimalla	Elena Esposito
Angela Grimminger	Reinhold Häb-Umbach	Barbara Hammer
Eyke Hüllermeier	Ilona Horwath (former member)	Friederike Kern
Stefan Kopp	Tobias Matzner	Axel-Cyrille Ngonga Ngoma
Katharina Rohlfing	Ingrid Scharlau	Carsten Schulte
Kirsten Thommes	Anna-Lisa Vollmer	Henning Wachsmuth
Petra Wagner	Britta Wrede	



<https://trr318.uni-paderborn.de/en/projects>





# TRR Conferences

TRR 318 hosted three international conferences:

The first conference, titled “**Explaining Machines,**” explored explainable artificial intelligence (XAI) from the perspective of the social sciences. In 2023, the second conference, “**Measuring Understanding,**” focused on evaluating and measuring understanding across various contexts. The third conference, “**Contextualizing Explanations,**” took place in 2025 and adopted an interdisciplinary approach to examine how explanations can be contextualized to enhance their relevance and empower users.



<https://trr318.uni-paderborn.de/en/conferences>





# Language for Talking about Explanatory Processes

**Explanation is at the heart of TRR 318 research, meaning that the terms used by TRR researchers in their daily work also need a quick explanation. An overview.**

## Co-construction

Co-construction refers to the interactive and iterative process in which partners jointly negotiate both the **explanandum** and the form of understanding for explanations. By sequentially building on, refining, and modifying each other's contributions, mutual participation is achieved, guided by **scaffolding** and **monitoring**. This process enables both partners to actively work towards a shared explanatory goal: while the explanation emerges on the microlevel of interaction, it is also crucially modulated on the macrolevel.

## Explainee

An explainee is the individual or group who receives an explanation. This term refers specifically to those who are intended to understand a concept or information being conveyed. In a classroom setting, for instance, when a teacher explains a concept, they are the **explainer**, and the students are the explainees, as they are the ones receiving and processing the information.

## Explainer

An explainer is the individual or entity responsible for delivering an explanation. This agent guides the process of clarification and understanding by presenting information, concepts, or ideas in a manner that is intended to enhance the understanding of the **explainee**. In a classroom setting, for instance, the teacher acts as the explainer, facilitating learning by conveying knowledge and answering questions.

## Explanandum

An explanandum is the entity, event, or phenomenon that is the focus of an explanation. It refers to what is being clarified or understood through the explanatory process. In the context of artificial intelligence, the explanandum might be the decision made by a credit scoring model, the classification of an image as a “cat,” or a recommendation for a specific movie based on a user's preferences.

## Monitoring

Monitoring is a multimodal process in which observed outcomes are compared to predicted results. Partners utilize speech, gestures, and nonverbal cues to track the progress of their joint tasks. The explainer evaluates the explainee's understanding to determine whether the explanation is effective or needs refinement. In turn, the **explainee** also monitors the **explainer**, gauging the appropriate level of detail required for a particular explanation.

## Scaffolding

In educational and developmental psychology literature, scaffolding is a concept that describes how an expert provides guidance to a learner by adjusting the level of assistance based on the learner's performance. In TRR research, this concept is adapted from the field of education and psychology to apply it to explanation in artificial intelligence encounters. Both partners in the interaction can scaffold each other, meaning they provide one another with the necessary information to collaboratively construct both the explanandum and the desired form of understanding. Alongside **monitoring**, scaffolding serves not only as a form of guidance but also as a means of supervision, facilitating the active participation of both partners in the learning process.

## Partner Model

A partner model is a key resource for “placing” explanations and encompasses knowledge and assumptions about the **explainee** regarding their role in the dialog, general characteristics, and specific attributes. It represents the mental model that a person has of another individual to whom they are explaining. This model includes prior knowledge, assumptions about the explainee's understanding, and their general traits. In the context of explainable artificial intelligence, the partner model is developed by the system through its interactions with the user, enabling tailored explanations that align with the user's needs and level of comprehension.

## Understanding

In the context of explainable AI, understanding refers to the relevance of information provided to users of AI systems. Unlike the current debate that focuses on simply giving “enough information,” the TRR's approach emphasizes that understanding should be tailored to what is important for the user.

The TRR researchers differentiate between two key concepts: enabledness and comprehension.

- **Enabledness** relates to how explanations help users make choices or take actions.
- **Comprehension** involves a deeper awareness that allows users to form a broader understanding of a phenomenon beyond what is immediately obvious.



# INTERVIEWS





# “We found new ways to better examine how explanations work”

*After four years of intensive research, TRR 318 speakers Professor Dr. Katharina Rohlfing and Professor Dr. Philipp Cimiano share key insights of the first funding phase: What have they learned about how artificial intelligence (AI) explains things? What were the challenges of bringing together researchers from different fields? And how has technological progress, such as large language models like ChatGPT, changed their work?*

## What were the most important new findings about explainability in AI?

### Katharina Rohlfing:

Important to our approach is the assumption that current explainable AI have a fundamental flaw: they treat explanations as a one-way street, that is, the machine explains, and the person listens. Our research contributed to the visibility of the explainee as the addressee of the explanation because in real life, understanding is a bidirectional matter: people talk to each other, ask questions, nod, look confused, or use gestures to show whether they understand something. That's why we developed a new framework that sees explaining as a two-way



© Stefan Sättele

process, like a conversation that unfolds in time and is shaped by the participants acting together. We call these systems that are designed according to our framework Social Explainable AI, or sXAI. They adapt their explanations to each individual in real time based on how the person responds or what they consider as relevant.

## How did you test whether this framework actually works?

### Katharina Rohlfing:

We closely studied real conversations, for example, how people explain things in everyday situations. What we saw is that even though most explanations start with a monologue, the explainee (person receiving an explanation) is usually actively involved: they ask follow-up questions, look puzzled, or signal their progress in understanding. That means explanation is often a dialogue, not a monologue.

Analyzing the explaining process more closely, we also looked at how people use language and gestures to signal what they understand and what needs to be further pursued. We identified certain patterns that show how people build understanding together. In addition, we analyzed ways of how they “scaffold” each other offering support step by step, like building a temporary structure to help someone climb. For example, one can first instruct how to do something but then

move on with explaining what not to do. Negative instructions could be a helpful scaffold.

**Philipp Cimiano:**

Our computational work was concerned with implementing our framework into AI systems. These systems respond to the person they are explaining something to. They consider three key aspects: cooperation, (how well the interaction works), social appropriateness (how suitably the system behaves), and understanding. The system SNAPE developed in project Ao1 is a good example. It is sensitive to how a person reacts and then adapts its explanation accordingly. So it doesn't give the same explanation to everyone, but customizes it depending on the situation.

**Did you develop new methods to study explanation more effectively?**

**Philipp Cimiano:**

Yes, we found new ways to better examine how explanations work. For instance, we developed new instruments to measure whether someone understood something from an explanation or was left confused.

And we didn't just limit our investigations to the lab environment. It was important to us to see how explainability plays out in everyday life, with different people in different situations. For example, we asked what kind of AI systems people use on their daily basis and whether they would like to have an explanation for their functions.

**Katharina Rohlfing:**

Our goal was to investigate how understanding develops – not just whether it happens. That's why we introduced a method where participants look back and describe their “aha moments” – those key turning points in their understanding. We put these moments into spotlight of our analysis. Another method was to design special workshops where people and AI work together to create explanations. These new methods are helping us to gain deeper insights into not only the process of explaining and understanding but also how to foster it and when explanation is helpful.

**What was especially challenging during the first funding phase?**

**Philipp Cimiano:**

The biggest challenge was bringing together people from very different disciplines like computer science, linguistics or psychology. Everyone has their own way of thinking and speaking. So first, we had to develop a shared language. Another huge challenge we faced came with the release of ChatGPT. This changed a lot of research concerned with technology development and opened up new possibilities to every user. That's why we quickly formed a working group to focus on new developments leading to new research projects.

**How well did interdisciplinary collaboration work?**

**Katharina Rohlfing:**

As a team, I am proud to say that we are strong in interdisciplinarity on many levels. Within individual projects, people from different fields work together. Thus, the projects have an interdisciplinary architecture. But we also work across projects, as it was the case for our first book on social XAI, which will be published



this year. On top of this, we work in groups pursuing hot topics that we consider relevant to our TRR, such as the group on LLMs.

Regular meetings, like our TRR conferences, writing retreats, and the so-called “Activity Afternoons”, also strengthened our collaboration. Of course, it’s not always easy to integrate new members into this established culture, but we’ve created formats that ease this process.

## What are the major challenges you see for the future?

### Philipp Cimiano:

Large language models, like ChatGPT, are powerful, but they also have limitations: they often don’t take the specific situation into account. They explain things, yes, but they don’t really understand who is asking or why. In the future, we’ll need systems that can adapt flexibly to the situation, systems that understand what’s relevant right now.

### Katharina Rohlfing:

We need to fundamentally shift how we think about explainability. It’s not enough for an output to be understandable. The systems need to create a context in which an interaction can be shaped by the users, so that people don’t just passively receive information but actively engage with it to come to a relevant understanding. This strengthens the collaboration between humans and AI and ensures that technology stays understandable but also useful.



Katharina Rohlfing is full Professor of Psycholinguistics at Paderborn University where she is heading the “Sprachspiellabor” for research on language development in children. She is involved in international and interdisciplinary projects in which she investigates multimodal social interaction, especially the process of scaffolding interaction partners and how robotic partners can achieve this. Her current research focuses on interactive adaptivity.



Philipp Cimiano is full Professor for Computer Science at Bielefeld University. He leads the Semantic Computing Group at the Cognitive Interaction Technology Center (CITEC). His fields of research include knowledge representation and engineering, natural language processing, knowledge acquisition and computational argumentation.

# “We Gained Greater Confidence as Researchers”

All early career researchers at TRR 318 are members of the research training group. Josephine Fisher works as a psycholinguist at Paderborn University and Roel Visser is a computer scientist at Bielefeld University. Both are completing their doctorates at TRR 318 and intend to graduate by the end of the year. In this interview, they reflect on their doctoral work in interdisciplinary Transregio projects and the mentoring they received from the RTG.

**You are both doctoral researchers at TRR 318, although you work in completely different disciplines and on different projects. Which are they?**

**Josephine Fisher:**

I am on project Ao1, which is called “Adaptive Explanation Generation”. We have researchers from linguistics like me, as well as from psychology and computer science. When people explain something, they usually incorporate the reactions of their conversation partners and adapt their explanations accordingly. We observe how people react towards each other during game explanations, and then we take this information and implement it into an explainable agent. My doctoral thesis is part of this project: my aim is to examine how explainees are involved into an explanation, a topic known as interactive adaptivity.

**Roel Visser:**

I am part of project CO1, which is called “Healthy distrust in explanations”. The project is a collaboration between the machine learning group in Bielefeld and the psychology group in Paderborn. We examine the interface between the people who use these systems and the systems themselves. The idea behind the project is that current machine learning and AI systems can be incorrect or produce errors, but they are still highly useful. Ideally, people are willing to use these systems, but do not put blind trust in them. If Netflix recommends the wrong movie to you, it doesn’t really matter that much. But if you are doing anything with self-driving cars or finance, this becomes more important. One aim of the project is to use explanations to enable humans to identify possible discrepancies in AI systems. This should help users to build healthy distrust towards these types of systems, rather than mere disagreement between machines and humans.

**Both of you started your doctoral research at Transregio in October 2021. Why did you want to join the doctoral program at TRR 318?**

**Roel Visser:**

I did my Bachelor and Master degrees in computer science in the Netherlands. I started my doctoral studies in Bielefeld in Barbara Hammer’s group. I liked the interdisciplinary aspect of the work. I was interested in these two different perspectives in the project – both the human-centric focus and the computational aspect. I think there is a gap in terms of people’s knowledge and expertise, but also in the available data. There is also a huge gap between the people designing these systems and people using them.

**Josephine Fisher:**

I did both my Bachelor and Master degrees at Paderborn University. I did a double major in English linguistics and English literature and cultural studies. When TRR launched, I was still working on my Master's thesis. I really liked working as a research assistant and analysing data and conducting studies. And I've also really enjoyed psycholinguistics as a specialization within linguistics. The main reason for me to join was that I could do full-time research and a lot of data analysis.

**How have you enjoyed doing your doctorate in this Transregio so far?**

**Josephine Fisher:**

I like working in an interdisciplinary team. Mine includes two other disciplines: psychology and computer linguistics. But you need all these different components. It helps to find your own position by thinking from different angles and perspectives. For example, when we were doing our data collection, everybody wanted something else from the studies. We had to implement another questionnaire for the psychology group. The computer linguists needed the data structured in a different way. So, you really had to think about how to get everything that everyone involved needed, thus broadening your own horizon.

**What experiences have you had during your research assign-**

**ments in Transregio projects? What challenges does working in an interdisciplinary team entail?**

**Roel Visser:**

Yes, sometimes different disciplines can get quite stuck into their own silo, also in linguistic terms. Then you need to keep re-evaluating, and make sure that you are all talking about the same thing and ultimately thinking about things from a common ground. It can be a benefit for everyone to have their own perspective, but if you are doing the work together, it has to be coordinated. You have to keep making sure that everyone is still on the same page.

**Josephine Fisher:**

And you must negotiate the terms of the conversation. It is also essential to have a positive mindset and appreciate the other disciplines. It is always important to discuss the hot topics and agree on the terms with each other, because the same terms are sometimes used in completely different ways in different disciplines. You learn a lot when you bring together these different disciplines.

**You are involved in different projects, but you know each other from your research training group (RTG) at TRR 318. Which role does this group play for each of you? Which kind of support was offered during your doctoral studies?**

**Josephine Fisher:**

The RTG is a graduate school. They want to support us in our academic careers, and provide concrete steps towards helping us achieve our goals. We started with workshops like: 'How to write an exposé?' or 'How to be a researcher?'. The RTG

supported us in many ways. For example, they put on a number of workshops that we could attend, where we also got to know other doctoral researchers. Having all those workshops and the opportunity to do writing retreats together really helped me. Last year I did a research stay for six weeks in Bangor in Wales. The RTG provided us a network that we could connect with.

**Roel Visser:**

I quite liked getting a bit of support from the workshops. It is also good to have support from outside of your project, beyond your supervisor and direct colleagues. It was interesting to hear the perspectives from other doctoral researchers: to learn about the kinds of problems that they run into with their research, or logistical or communication issues. It can be difficult as the new kid on the block coming into the research project. So, it is quite helpful to get this support to help you hit the ground running.

## **What is the value of the interdisciplinary network of early career researchers in the RTG?**

**Josephine Fisher:**

Usually, when you start at a new institution, you're the new person. For us in the RTG, we were all new together. Being in the same boat helped a lot. We were connecting with each other. I probably wouldn't have met so many different people if I hadn't been in TRR. I liked talking to all the different people and finding out what their worries are and how their work is going. We were building team spirit. Being a TRR member is now part of who I am in my professional life.

**Roel Visser:**

During doctoral work, you can get stuck on an island, so to speak, and get really focused on just your own research projects. In addition to this, you can also be on an island within your own research group, even with your professor and the other doctoral researchers in that group. I like how the RTG helps you to get off your own island.

## **How did you develop on a personal level during your doctoral studies at TRR 318? What's on the horizon after your doctorate?**

**Josephine Fisher:**

Thinking back to 2021 and comparing myself to today: I gained a lot more confidence as a researcher. I know that there are still a lot of things I need to learn, but you come to understand that this is part of being a researcher. You can't know everything. I'm also more confident in connecting with people. I learned to put myself out there and be confident that what you've got to say about this topic matters, and that you impact the research community. I also organized the second TRR conference and that was a huge learning experience. In terms of the future, I am hoping that we secure a second funding phase.

**Roel Visser:**

I enjoyed gaining experience in conducting research, writing papers, and learning a lot about the research process. This helped build confidence for me as well. I also liked getting more and more immersed in the topics, which enables you to gain deeper expertise over time. In terms of the future, I am not yet sure of which direction I want to go into. At the moment, I am open towards going into academia or industry.







# RTG Brings Together Early Career Researchers from Many Disciplines

TRR 318's graduate school provides a structured training program for doctoral researchers and post-doctoral researchers. Ricarda Kurock is the coordinator of the Research Training Group (RTG). In the following interview, she speaks about the offerings, goals, and success of the program. A key element for her is the cooperative, interdisciplinary approach.

by Jana Haver

## What is the goal of the Transregio's Graduate School?

### **Ricarda Kurock:**

Our mission is the further education of researchers who are associated with TRR 318 or who are employed there. We support doctoral and post-doctoral researchers in their work during early career stages by providing them, on the one hand, with deep insight into the various disciplines, and on the other hand, by helping researchers develop the skills to participate productively in interdisciplinary collaboration.

## How does this interdisciplinarity manifest in the RTG?

### **Ricarda Kurock:**

Interdisciplinary collaboration is at the heart of everything we do; it's not just an add-on. Our researchers come from a wide range of disciplines—from linguistics

and psychology to computer science—and collaborate closely. Interdisciplinarity can be extremely productive when, for instance, theories or methods from different disciplinary perspectives are brought together. Indeed, interdisciplinarity offers many opportunities, but it also poses challenges. For example, if two people are talking about a tree and one person is thinking of a pine tree and the other of a broadleaf tree, but this is not communicated, it can quickly lead to misunderstandings. This sometimes happens when different disciplines come together. That's why it's so important for us to empower early career researchers to successfully navigate such processes.

## How is the Transregio's RTG structured and what services are available?

### **Ricarda Kurock:**

An essential part is the workshop program where personal and professional skills are honed. We also offer various events and counselling services. Our writing day events and retreats, which provide a shared space for academic writing structured by our writing consultant, Andrea Karsten, are particularly popular. We are also committed to providing individual support to early career researchers at TRR. We hold annual development meetings with our doctoral researchers to review their contributions to the research being done at TRR, but also to reflect on their own personal development. Beyond this, we also offer mentoring and networking opportunities. Another important function of the RTG is to provide financial and social support, such as for research trips abroad.



## **You mentioned networks that early career researchers can build and use. What exactly do you mean by that?**

### **Ricarda Kurock:**

The aim of the RTG is to foster a close-knit community of research practitioners. The focus here is on networking—with each other, but also externally. Early career researchers from different disciplines and projects can come together and exchange ideas through the RTG. But there is also external input. We hold regular events where we invite various people to speak on specific topics, for instance, former academics who are now working outside of academia. With these kinds of offerings, we want to highlight different career paths. Talks on experiences abroad are also popular. For example, we invited an assistant professor to discuss her experiences working in the US and answer questions about it.

## **To what extent can early career researchers pursue experiences abroad themselves?**

### **Ricarda Kurock:**

We fund research stays abroad. So far, around 15 researchers have already done a research stay abroad, at locations ranging from Copenhagen to Australia. We also support doctoral researchers who want to visit the Transregio: we have hosted guests from Germany, Italy, Poland, and the Netherlands.

## **How successful is the RTG? How many researchers have been educated there?**

### **Ricarda Kurock:**

Most of the doctoral students are currently in the final spurt of their doctorate; we cannot yet say who will finish when. We continuously evaluate our program in order to provide the best possible offerings. Our writing retreats are very popular: a group of early careers researchers go to a hotel where they can get support on the writing process and have the opportunity to work on their individual projects. The structure and the shared setting give participants both autonomy and social connection, the combination of which is very motivating. Here researchers are not only supported in their writing, but also talk to each other and share their experiences. There are writing retreats on the calendar in the coming months.

## **What are the skills and abilities that doctoral researchers get from the RTG, above and beyond what they're learning from the content of their own projects?**

### **Ricarda Kurock:**

The greatest strength of the RTG really lies in interdisciplinary exchange. Different perspectives and methodological approaches come together. Concepts are discussed and further developed in ways that would not be possible without exposure to this diversity of viewpoints. The methodological approaches are also often very different. RTG members are constantly encouraged to communicate with each other, to refine their own point of view, to develop a sense of openness towards other perspectives, and to make compromises. I think that this is of great value. And because we communicate in English, this is of course also a skill for the international job market.

## How can prospective students become part of the RTG?

### **Ricarda Kurock:**

Researchers who are employed in TRR 318 and are in an early funding phase, i.e. doctoral and post-doctoral researchers, are automatically part of the RTG. External doctoral researchers or students can also apply if they are working on a TRR-related topic.

## What's up next for the RTG in the next phase of funding?

### **Ricarda Kurock:**

We hope that the RTG and TRR 318 as a whole will be able to secure a second round of funding. We gained a lot of experience at the beginning and were able to optimize some aspects quite well. This is why we would like to continue and contribute the knowledge we have gained and further expand the network we have already established.

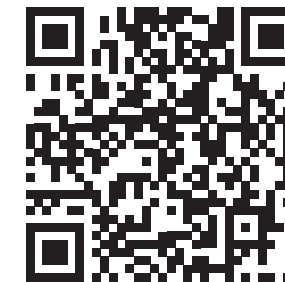
## What does the RTG and your work there mean to you personally?

### **Ricarda Kurock:**

I've been at the RTG for approximately two years. I am currently training to become a systemic counselor. Counseling always requires a bit of relationship work, so it took some time to get to know the members and understand their needs. But everything is going very well now. I really enjoy my work and I think it's great to see how our service offerings are being used and how the researchers are developing. I am delighted to be able to support them.



© Susanne Freitag



<https://trr318.uni-paderborn.de/en/research-training-group>

The background is a light yellow-green color. It features several overlapping circles of different sizes and shades of green and grey. On the left side, there is a blue speech bubble shape. Inside this bubble, the text "SYNTHESIS GROUPS" is written in a bold, blue, sans-serif font.

# SYNTHESIS GROUPS

# TRR 318 Synthesis Groups in the First Funding Phase

TRR 318 synthesis groups are organized around emerging topics that spark interest and are relevant within the broader scope of the consortium's research. A call for ideas was made to all TRR members, inviting them to propose innovative topics worth exploring. Through a collaborative, bottom-up process, these ideas were evaluated by the entire TRR community, allowing members to express their interest and commitment by joining a synthesis group.

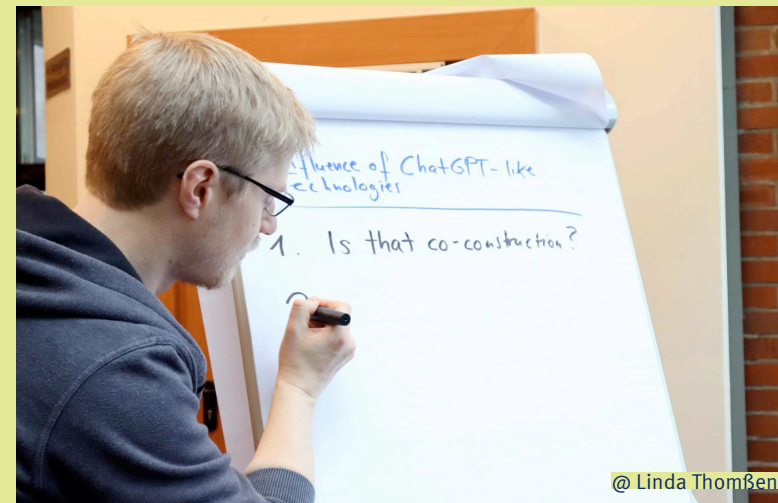
The synthesis groups are interdisciplinary, bringing together diverse expertise to address key issues in the field, including:

- **“Architecture(s) of Co-Constructive XAI Systems”**: Aims to develop a unified computational architecture that allows for the implementation of co-constructive explanatory systems. The group developed an architecture based on the MAPE-K framework.
- **“Impact of Generative AI”**: Investigates how the recent advent of generative AI technologies, such as large language models (LLMs), are impacting the goals and opportunities of TRR 318. The team is working on a first prototype of a co-constructive LLM that can undertake explanatory dialogs with humans.
- **“Processuality in Interaction”**: Investigates how interactions between people dynamically develop and evolve. The focus is on the often-unconscious ways in which conversation partners adapt their communication to each other.

- **“Theory Development”**: All project teams contribute to the development of theory. To illustrate and illuminate the work and findings from individual projects, all Transregio members contribute theoretical building blocks. In addition to advancing the underlying theory, these blocks also present empirical results, software/architectures, and ontological developments.

- **“Understanding”**: Deals with the concept of understanding and its significance in the context of explainable artificial intelligence (XAI). Its focus is on the distinction between two main aspects of understanding that occur during the explanation process: comprehension (“knowing that”) and enabledness (“knowing how”).

- **“Unpacking Concepts”**: Unpacks core conceptualizations and their implications in XAI research. Its approach is to analytically examine publications and other documents frequently cited in the XAI community to identify conceptual alternatives that direct research in specific ways. The goal here is for researchers to develop a greater awareness of the implications, historical burdens, and context of related scientific concepts.



# The Goal of Explaining is Understanding

**The “Understanding” synthesis group is one of six synthesis groups at TRR 318. Some 20 researchers from a variety of disciplines and Transregio projects have joined forces to develop a concept of understanding and elucidate its significance in the context of explainable artificial intelligence.**

by Jana Haver

Transregio synthesis groups were created as cross-sectional groups. The goal of these synthesis groups is to discuss key aspects of research from a higher-level perspective and enable the individual projects to benefit from this work. The various synthesis groups were created based on personal interests—researchers were able to make multiple selections.

Professor Dr. Heike M. Buhl is part of the three-pronged organizational team for the “Understanding” synthesis group. “The goal of giving an explanation, in general terms, is that people ultimately come to understand something.” And this is precisely what the researchers in this synthesis group have been investigating during the past years.

## Different Kinds of Understanding

Understanding can take on a variety of forms. The focus in this synthesis group is on two types of understanding: comprehension (understanding that) and enabledness, the ability to do something (understanding how). “Someone might want to understand something in order to perform an action,” explains Heike Buhl. “So if I say, ‘How do I turn on the light?’, the explanation my counterpart might give is, ‘Press the switch’. Then I am able to act and can turn on the light.

And If I ask my interlocutor ‘how does that work,’ a possible explanation might be something like ‘by pressing the switch, you toggle a lever, which establishes a connection, thus allowing electricity to flow and the light to turn on,’” as Heike Buhl explains. Now, the person is able to comprehend the electrical system itself.

There is also an important distinction between shallow and deep understanding. Depending on the situation, a simple explanation may suffice, or a more in-depth explanation may be required. “In daily life, shallow understanding often suffices,” says Buhl. “One does not always need to know exactly how something works—it’s enough to know that the light goes on when you press the light switch.” But there are also situations in which deep understanding is needed, whether for scientific work or technical design. In order to design a lighting system, for example, one would need to have deep ‘enabledness’ type of understanding of the ‘how’ behind the system. When artificial intelligence is used in everyday situations of limited consequence, shallow understanding is probably adequate for most users.

## Working in Smaller Groups within the Synthesis Group

After reaching a general agreement on the concept, the “Understanding” synthesis group discussed the various forms of understanding and worked these out in greater detail. To do this, they split off into smaller groups to consider and analyze the topic using predefined research questions.

The goal of the synthesis group was then to come up with a jointly agreed upon definition of understanding and its types, and to examine the significance and dynamics of understanding in everyday life. “We looked at the process of explanation and asked questions like ‘What happens when we explain?’, ‘How do we



move from shallow to deep understanding?’, and ‘What is the starting point of an explanation?’” says Heike Buhl.

## Interdisciplinary and Inter-Project Cooperation

Researchers from the “Understanding” synthesis group organized themselves into individual groups according to their interests. As Heike Buhl asserts: “I thought it was great that the groups were not only diverse and interdisciplinary in terms of their project affiliations and disciplines, but that they also included researchers with different levels of experience.” Buhl sees a significant benefit from the additional opportunities for collaboration within the synthesis group, adding: “Members hail from all areas of the TRR.” This leads in a lively exchange of new ideas. In addition to this, researchers bring their different backgrounds and the goals of their individual TRR projects to bear. “Of course, everyone asks themselves: ‘What can a synthesis group offer me for my project, and what findings can I use myself?’” says Heike Buhl. “In my opinion, this was a very cooperative process, as it brought together researchers who would not normally have collaborated on their everyday work.”

The smaller groups within the synthesis group then worked together on pre-defined guiding questions. Their results were regularly discussed and debated within the larger synthesis group as a whole. “That was a very complex process,” explains Heike Buhl. “The process went largely from the bottom up, so that everyone could contribute to the structure itself.”

## Everyone Benefits from Findings

Heike Buhl herself was in one of the smaller groups—one that was very diverse. “The group consisted of computer scientists, a mathematician, a linguist, a computer linguist, and me as a psychologist.” For Professor Buhl, looking at the same topic from so many different angles is extraordinarily advantageous. “We were all trying to answer the same question: ‘How can one describe the starting point and development of understanding?’” As Buhl explains: “A computer scientist is very good at systematically describing processes, while the linguist elaborates the development of expertise.” For her part as a psychologist, Buhl was able to contribute her expertise in learning processes and different forms of knowledge. “This is great. We share a common interest in the topic and are working on it—together,” says Heike Buhl.

The article “Forms of Understanding of XAI-Explanations” arose out of the work done by this synthesis group. Heike Buhl is pleased with what they accomplished: “our findings have been very well received,” she says, and “our conceptualization has been referenced extensively in research proposals.” The psychologist hopes to continue working with the synthesis group, further elaborating and building upon their initial findings, for instance, by investigating how understanding can be measured.



The background is a solid light green color. It features several overlapping circles and speech bubbles in shades of grey and lime green. On the left side, there is a blue speech bubble icon. A white rectangular banner with rounded ends is positioned horizontally across the upper left, containing the text "PUBLIC ENGAGEMENT" in blue, bold, sans-serif capital letters.

# **PUBLIC ENGAGEMENT**

# Engagement Beyond the Lab

At TRR 318, researchers actively interact with the public through various outreach initiatives and by collecting data in real-world contexts. This section showcases how the research team interfaces with different audiences and fosters discussions about artificial intelligence.

## Hospital Study

The linguists from projects A01 and A04 analyzed interactions between doctors and patients at a children's hospital to determine who is responsible for initiating each phase of an explanation. The results show that, in this setting, monologues from the doctor tend to dominate explanations, while dialog-based phases are primarily initiated by the patients or their caregivers. This transdisciplinary finding has provided the research teams with new insights into the interactive responsibilities of the respective roles in explanations (Explainer – doctor, Explainee – patient/caregiver). However, the findings also raise the question of how patients can be empowered to actively participate in the explanatory dialog.



© Mike-Denis Müller

## Digital Diaries of Lay Users

In the research conducted for project B03, the team used digital diaries to investigate the everyday use of explainable AI (XAI). They found that lay users can become expert users for practical purposes in their daily activities. The analysis revealed that when using AI, most users rely on sense-making processes to fulfill specific explanatory needs, often rendering formal explanations unnecessary. Instead of seeking structured explanations, users tended to require AI explanations that were highly contextual and situational, shaped by their roles, prior knowledge, and the perceived stakes in their interactions with AI systems. In follow-up group discussions, participants shared and debated the experiences they had recorded in their digital diaries, which captured the complex arrangement of AI encounters and explanatory needs in everyday life.



## Study at Police Departments

This study, conducted by project Bo1, focused on explainability in the context of predictive policing in German-speaking regions. Using qualitative social research methods, including semi-structured expert interviews and ethnographic field observations, the research aimed to explore how machine learning (ML) predictions—specifically, predictions of burglary risk in certain residential areas—are communicated across different departments. It identified points in the chain of communication where understanding difficulties, questions, and explanations arise. In a pilot project with three police departments, the Bo1 team tested an XAI application, providing it to so-called “analysts.” Initial results indicate that the way explanations for machine learning predictions are requested and given depends on the organizational context. This includes factors like the level of formality, the methods of communication used, and the various roles involved in the communication process. Furthermore, the research suggests that implementing XAI methods not only addresses explainability issues but also integrates into the existing system of explanatory communication within the police organization.



© Michael Lenke

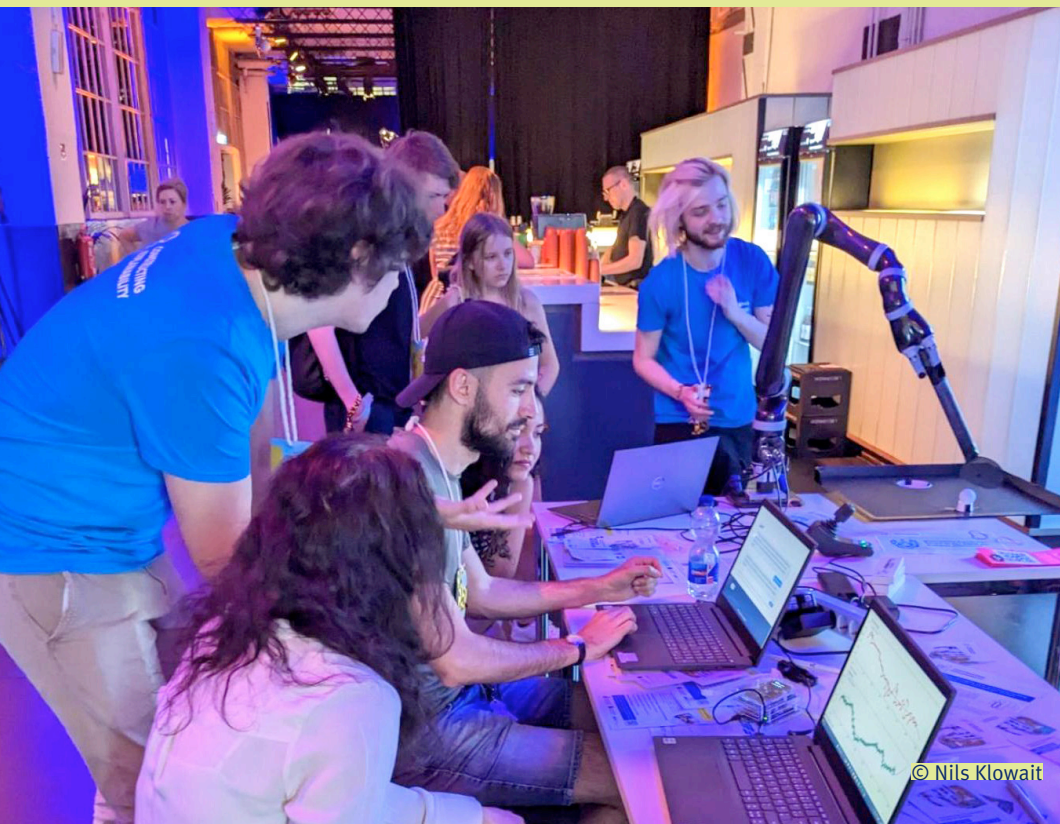
## Co-Construction Workshops

Members of the general public had the opportunity to experience AI systems firsthand at the co-construction workshops organized by project Ö. Participants learned about how AI works and how data impacts AI systems. They also discussed and debated the implications of using AI systems. The workshops fostered many fruitful discussions about AI, touching on topics such as ethics and biases, the limitations of AI decisions, and the question of responsibility for those decisions. While these issues could have been addressed from a “top-down” approach with lectures from AI experts, the members of project Ö instead positioned themselves as curators of an environment where the individual voices of participants came through.

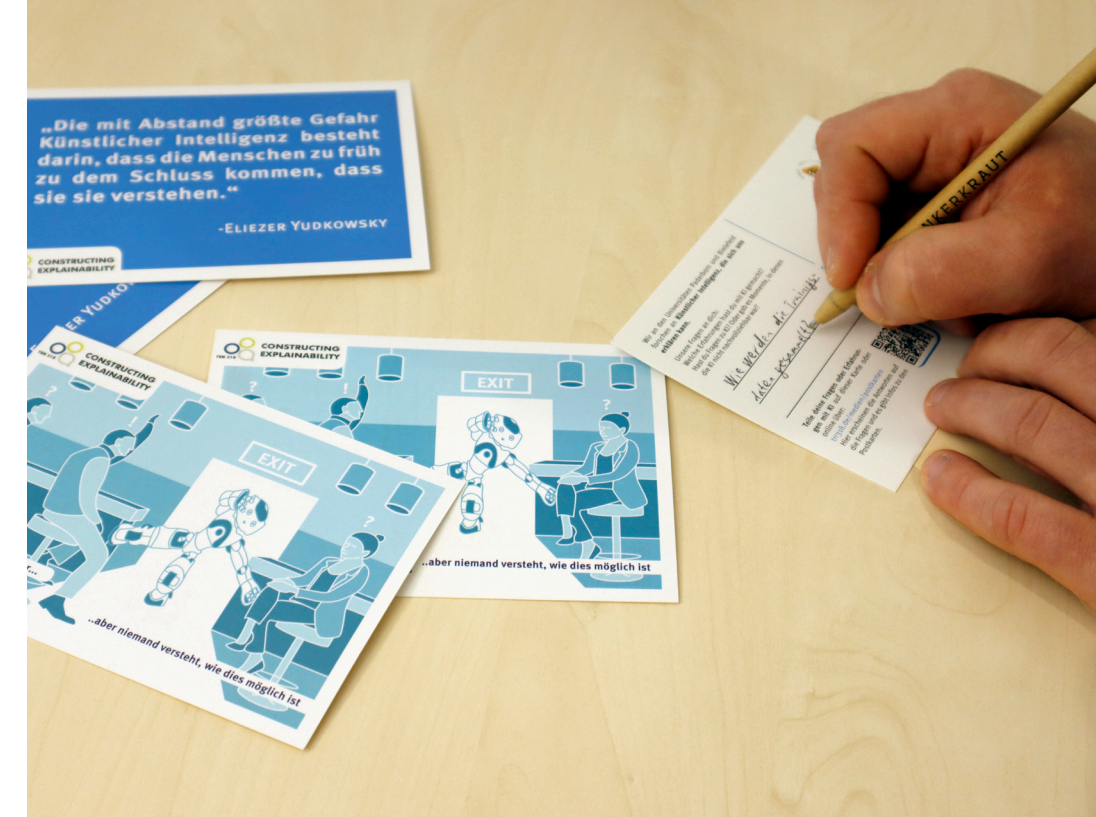


## Presentations to the Public

TRR researchers had the opportunity to showcase their findings to the public at several fairs and events. A highlight was the re:publica 2023 festival, where the teams from projects Bo5 and Ö brought their systems to Berlin. During the event, they demonstrated how their AI technologies work and discussed the implications of these systems, fostering a lively exchange of ideas with attendees. The interactive nature of these presentations allowed visitors to experience the AI systems firsthand, which helped to demystify the technology and demonstrate its practical applications.



© Nils Klowait



© Linda Thomßen

## Engaging the Public through Postcards

Through a postcard campaign, the members of project Ö invited members of the local community to engage in a dialog about AI. The postcards featured two different designs and were handed out at pubs throughout Bielefeld and Paderborn to reach individuals who might not otherwise approach TRR experts with their questions. The printed QR codes on the postcards linked to the TRR 318 website, where participants could submit their comments and questions about AI. The postcards could also be mailed directly to the project team with a question. The experts at TRR 318 then provided detailed answers, which are now available on the project website at [www.trr318.de/en/media/questions-and-answers](http://www.trr318.de/en/media/questions-and-answers).

# Making Better, More Explainable Medical Diagnoses with AI

**Can artificial intelligence work side by side with doctors to help them make better diagnoses? A research team of computer scientists in project Co5, called “Creating Explanations in Collaborative Human-Machine Knowledge Exploration,” is dedicated to answering this question. They are working on an interactive AI system that accompanies doctors as they make a medical diagnosis by reviewing and evaluating assumptions in dialog with them.**

by Silke Tornede

Artificial intelligence (AI) is becoming increasingly important in medicine: intelligent systems can sort through large volumes of data in seconds flat, analyze X-ray or MRI images, identify abnormal findings, and even make diagnoses. But how does AI arrive at its conclusions? This is where project Co5 at the “Constructing Explainability” TRR 318 at Bielefeld University and Paderborn University comes in. As Professor Dr.-Ing. Stefan Kopp explains, the challenge here is not only about providing “diagnostic support, but also reasoning support.” The goal is not for the AI to provide ready-made medical recommendations per se, but rather to assist medical professionals in their decision-making process by asking questions and discussing hypotheses. “Kind of like a lane-departure assistance system in driving, this AI too would ensure that the doctor does not overlook anything important during the diagnostic process and would provide guidance on what else might need to be considered,” says Stefan Kopp, who is working on this basic research together with Professor Dr. Philipp Cimiano and two doctoral researchers.

At the heart of this system is the ASCODI system, a prototype developed and refined by computer scientists and cognitive scientists to enable humans and machines to work together as a team and arrive at a good diagnosis—step by step. “And this means not only being as accurate as possible, but also understandable and justifiable,” says Kopp.

To show how this works, doctoral student Dominik Battefeld turns on his screen and slips into the role of a doctor using the ‘assistive co-constructive differential diagnosis system,’ or ASCODI for short. He clicks on the button ‘Start Diagnosis’ and is presented with a simulated patient: female, 32 years old, presenting with a severe cough. Dominik Battefeld checks additional information on the dashboard and then asks the AI: “What should I do next?” ASCODI reminds him to check for fever and vomiting and provides him with an assessment of how likely or unlikely illness X or Y is. “It’s like a virtual colleague with whom I can discuss the case,” says Battefeld. The smart assistance system provides answers, but also intervenes proactively, asking questions and setting red lines that should not be crossed where there is uncertainty or doubt. For example, the system warns that, at this point in the diagnostic process, it is not warranted to rule out the hypothesis that Dominik Battefeld is trying to delete.

## Making AI Understand How Doctors Think

In order for ASCODI to be able to do all of this, the AI system was first uploaded with medical information, which was then used to train it to associate symptoms with diseases. This included a dataset with 1.3 million simulated patients. But the Bielefeld cognitive scientists are taking this a step further: their model is meant to understand how doctors think and how they make diagnoses. “This doesn’t yet exist,” emphasizes Stefan Kopp. Medical decisions are complex, and every doc-



tor has their own strategy. The researchers distinguish between two broad types of diagnostic approaches: those who develop hypotheses quickly and intuitively based on knowledge and experience, and the more analytical types, who gather a great deal of information and work more slowly. ASCODI should recognize the doctor's process—and help counteract cognitive biases. Maybe the doctor has a lot of expertise in one area, which could run the risk of overlooking other possibilities. Doctors might commit to a diagnosis too early, only paying attention to what supports their assumptions. In practice, such misjudgments occur again and again, especially when working under time pressure, as Kopp explains. “The error rate in diagnoses is estimated to be up to ten percent.”

## **Can ASCODI Reduce the Error Rate in Diagnostics?**

The researchers regularly discuss their findings with medical professionals from clinics throughout Germany. Philipp Cimiano, for instance, conducted a study to investigate how doctors respond to AI recommendations and how they rate these suggestions in their professional opinion. “We are developing approaches that allow the diagnoses suggested by AI to be explained and justified to the treating physicians,” Cimiano says. Previous findings have shown that conventional approaches to explainable AI often fall short. “Doctors need well-reasoned justifications that present the facts and draw statistical conclusions from them. However, the rationales should not be too complex, as doctors have limited time to consider explanations from AI systems.” Cooperation partners on this project include the Epilepsy Center in Bochum and the University Clinic for Epileptology at the Mara Hospital in Bielefeld-Bethel. Epileptology is well suited to this research because the investigation of epileptic seizures is highly complex. Dr. med. Christian Brandt, who heads the Bethel Epilepsy Center and is a professor

of Epileptology at the University Hospital, sees promising opportunities for this technology to help make diagnostics faster and more precise: “However, the risk is that the AI becomes an opaque black box.” This is exactly what ASCODI is designed to prevent.

Further studies are needed to demonstrate whether the prototype meets these expectations. How does working with an AI affect doctors' behavior? Does this collaboration actually work in practice? What are the limits of this? Does the technology maybe even lead users to rely on it excessively—handing over too much responsibility to the machine? More testing will be carried out in the coming year. Another idea is that the system could be used in medical school or as a training tool. In any case, for the researchers it is a scientifically exciting project with social relevance. “Having a mathematical model of how and why doctors arrive at different diagnoses is a scientific challenge and it has added value if it works,” says Dominik Battefeld. Stefan Kopp also sees great potential in this basic research: “Using technology to support doctors who are under a lot of pressure and help reduce misdiagnoses is an exciting perspective for this promising application.”







# OUTLOOK



# A Forward-Looking Perspective on Future Research Challenges

A key challenge for XAI research, and especially for TRR 318, is how to make an explanation relevant for a particular user. “Because AI is pervasive in our lives, there is an urgent need to move towards a truly social and interactive view of explainability,” says TRR 318 speaker Prof. Dr. Katharina Rohlfing. “We can foster users’ understanding with an epistemic environment, for which systems have to be able to co-construct explanations with users.”

Looking ahead to the potential second funding period starting in 2026, TRR 318 researchers want to continue developing co-constructive explainable as well as explaining systems, thereby advancing the paradigm of social XAI (sXAI). Within this paradigm, XAI systems engage the human explainee and work together to arrive at a suitable explanation.

Additionally, the TRR team will work from the standpoint that suitable explanations can only be developed by considering the specific circumstances under which users engage in the explanation process. According to Katharina Rohlfing, “An explanation can be considered as relevant when assessed with respect to the context in which the explanation process takes place. The process includes not only the user but also the task, the environment, previous interactions, and much more.”

To achieve this vision, TRR projects will continue to focus on interactive settings where users can take part in shaping the process of explanation. “In order to maximize relevance, the explanatory algorithms need to be able to proactively determine relevance in a particular context,” explains Katharina Rohlfing. “This requires a framework and context

modeling that are more explicit, detailed, and interaction oriented than have previously been considered in the field of XAI.”

The emergence of powerful large language models (LLMs) only underscores the importance of TRR’s focus for the second funding period. “Because LLMs allow users to interact with them, they appear co-constructive,” says TRR 318 deputy speaker Prof. Dr. Philipp Cimiano. “For the second funding period, we have thoroughly analyzed whether current LLMs are capable of co-construction. We have determined that while they are interactive, they are not co-constructive; overall, they lack sensitivity and awareness for a context that is not only brought into the interaction but emerges from it.”

In the second funding period, the TRR researchers aim to utilize context in their models to identify and predict variables that will help ensure that a relevant explanation is created. “This research line has the potential to make significant contributions across disciplines. Our context-sensitive XAI systems will make predictions, monitor relevant factors, plan a successful explanation process, and explicate contextual variables. Subsequent interaction behaviors will, in turn, modify the emerging context,” says Philipp Cimiano.

Based on this continuous context-awareness, Transregio’s XAI systems will be able to provide suitable explanations and proactively scaffold users’ understanding. “By embracing LLMs in many projects, we are considering how context needs to be encoded so as to be successfully integrated into an LLM. In the long-term, these systems will empower users by providing information flexibly and in accordance with users’ emerging need for understanding – and questioning – how these systems arrived at the decisions they did.”

**The researchers at TRR 318 are working to develop context-sensitive, co-constructive systems that proactively build and adapt context incrementally, ensuring relevance and strengthening human-AI collaboration.**

# For the Latest News

The latest developments in TRR 318 research are published regularly on our website and social media accounts:

- Website:  
[www.trr318.de/en](http://www.trr318.de/en)
- LinkedIn:  
<https://www.linkedin.com/company/trr-318-constructing-explainability/>
- Instagram:  
[https://www.instagram.com/sfb\\_trr318/](https://www.instagram.com/sfb_trr318/)
- Newsletter:  
<https://trr318.uni-paderborn.de/en/media/newsletter>
- Podcast:  
<https://trr318.uni-paderborn.de/en/media/explaining-explainability>