ANNUAL REPORT

The Hybrid Intelligence Centre - 2025

Augmenting Human Intellect

Contents

01 Executive Summary

02 Highlights

03 Internal Collaborations

04 The Organisation

05 Education & Training

06 Finance

07 External Outreach

08 International Conference

09 Ambitions Year 6

10 PhD/PD Projects

01

Executive Summary

The year 2025 was the 6th full year of the Hybrid Intelligence centre.

2025 was a year of growth, consolidation, and strategic positioning for the HI consortium. The management team, strengthened by Erman Acar (UvA) and Nazmiye Ozkan (Hands4Grants) launched the recruitment of the 3rd research cohort, strengthened the collaboration with ALGOSOC, and with the help of Shenghui Wang (UT), organised a second round of engineering funding. A dedicated strategy meeting marked an important step toward institutionalizing the Science of HI and securing continuity beyond the current grant period, with the creation of a curated joint GitHub repository as its first tangible outcome.

The consortium maintained a strong scientific output, with over 90 publications, two international workshops, three SIKS courses, and a record attendance the HHAI 2025 conference. Members received major recognitions, including an ACM Fellowship (Maarten de Rijke), a EurAI Distinguished Service Award (Frank van Harmelen), an NWO Vidi grant (Sara Magliacane), and an EU Horizon grant (Davide Grossi). Eleven PhD candidates: Bram Renting, Jonne Maas, Lea Krause, Maria Heuss, Merle Reimann, Michiel van der Meer, Nicole Orzan, Selene Baez Santamaria, Tiffany Matej, Urja Khurana, Wijnand van Woerkom defended their theses, and several researchers advanced to senior academic positions. We thank them all for their contribution to the HI consortium and wish them well in their next career steps.

We are proud of our efforts to have an HI impact on society. Davide Dell' Anna and pInar Yolum with their students developed and applied the VIVE tool for the 510 Red Cross Initiative by extending the Axies tool of Enrico Liscio. Kim Baraka successfully launched our Artist-in-Residence programme. Wendy Aartsen with her colleague Jos Mitsunaga established the company HI-Collective. Finally, we bolstered our public outreach by the release of several consortium podcast episodes, featuring our PhDs and postdocs discussing HI themes with moderator Jos Mitsunaga. These activities demonstrate the growing reach of Hybrid Intelligence beyond academia. Still, there is a long way to go to achieve our mission to have AI augmenting, not replacing humans in our society. Therefore, we are quite pleased with the help of our newly installed Societal Advisory Board that has representatives from government (Amsterdam and Rotterdam), industry (KPN, ASML, Exin, Improven), consulting (Deloitte, KPMG) and healthcare (Universiteit Maastricht). The board encouraged us to organize outreach events for companies and organizations, as well as to strive to influence governmental bodies in changing policies and legislation that steer the use of AI in society to the HI way!

Overall, we look back at 2025 as a year in which we strengthened the consortium's scientific foundations, organisational sustainability, and societal impact.

Frank van Harmelen has been honoured with the **EurAI Distinguished Service Award** at the European Conference on Artificial Intelligence (ECAI), held this year in Bologna.

Feline Lindeboom, PhD candidate at the University of Groningen, has been selected as one of the **new Faces of Science by the Royal Netherlands Academy of Arts and Sciences (KNAW)**.

Sara Magliacane receives **NWO Vidi grant** for research on understandable AI

Maarten de Rijke was awarded **Prestigious ERC Grant** and **becomes ACM Fellow**

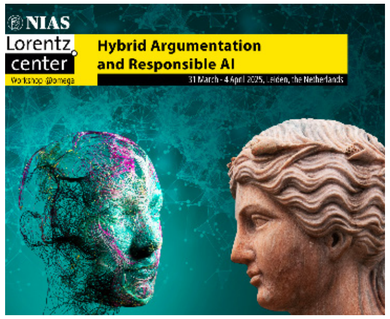
Davide Grossi has received funding from **ARIA** for HI-Related Project on AI Safety

Guidelines for Guidelines held on May 9 in Delft

In this workshop, participants investigated how effective guidelines could bridge different practices and foster mutual understanding: guidelines for guidelines.

Hybrid Argumentation and Responsible AI held on March 31- April 4 in Leiden

The event brings together around 50 invited participants from diverse disciplines—philosophy, linguistics, computer science, and hybrid AI—who will spend the week exploring how argumentation can support responsible interaction between humans and intelligent systems.



HI Researchers Selected for AmSIA Pre-Finals

We are proud to announce that **Siddharth Mehrotra and Maryam Alimardani** from the Hybrid Intelligence consortium have been selected as pre-finalists in the Amsterdam Science & Innovation Award (AmSIA) 2025.

We published over 90 papers in a variety of international venues and journals

Record-Breaking Interest in HHAI 2025 Conference: held on June 9-13 in Italy

HI methods at the 510 Red Cross Initiative:

How can we better understand what matters most to people during a crisis? MSc student Raoul Brigola from Utrecht University, in collaboration with The Netherlands Red Cross initiative 510 and under the guidance of HI members Davide Dall'Anna and Pinar Yolum, explored this question using Hybrid Intelligence.

Frans Oliehoek has been appointed Full Professor

Frans Oliehoek has been appointed Full Professor at TU Delft, holding the Chair of Interactive Learning and Decision Making within the Faculty of Electrical Engineering, Mathematics, and Computer Science

We are happy to announce that most of our 1st cohort PhDs have successfully graduated and most of our 3rd cohort PhDs and PDs are onboarded

Highlights

02

Challenges

Collaboration & Synergy (Cosy)

Leaders: Erman Acar, Roxana Radulescu

Progress in 2025: CoSy meetings were re-launched in May 2025. We identified the common themes and research question mapping across the newly joined 3rd cohort projects. We contributed to the Collaborative research line in the HI 5-year overview project, directly offering us insights into the progress and extent to which the CoSy research questions have been addressed and evolved.

Ambitions in 2026: CoSy considers proposing a workshop for HHAI 2027. The challenge team would like to find a joint project (e.g., white paper, opinion piece, blog post series) to involve all the participating projects, with their different perspectives and methods.

Dialogue

Leaders: Catharine Oertel, Piek Vossen

Progress in 2025: During 2025, the individual PhD projects made clear progress towards the scientific goals of the Dialogue Cluster. At the same time, the cluster has not yet achieved the level of convergence we would like to see across PhD and postdoctoral projects in relation to a shared research agenda.

Earlier in the year, we made several attempts to encourage PhD candidates and postdocs to self-organise around a joint demo, challenge, or shared research output. These attempts showed that bottom-up convergence is difficult to achieve without more structured support. PhD candidates are understandably focused on the publication and progress requirements of their own projects, and cluster-level activities can easily be perceived as additional work. Moreover, not all members immediately see how their individual projects connect to one another or to the larger objectives of the Dialogue Cluster.

This highlighted an important challenge: while the individual projects are scientifically strong, their contribution to the broader mission and vision of the cluster needs to be made more explicit. We therefore decided to focus on two strategies.

First, we will strengthen the Bubble Board engineering effort as a practical instrument for integration. The aim is to support PhD candidates and postdocs in connecting their individual research to a shared technical and conceptual framework.

Second, we have started one-on-one meetings with cluster members. These meetings are intended to clarify the connections between individual projects and the cluster's overarching goals, while also identifying how the cluster can better support each researcher's own trajectory.

Ambitions in 2026: Looking ahead to 2026, our aim is to create stronger alignment between individual project progress and the collective mission of the Dialogue Cluster, so that contributions to the broader cluster become mutually beneficial rather than an additional burden.

Overall, we maintain our ambition is to advance dialogue research as a foundation for Hybrid Intelligence systems that support long-term, meaningful human-agent interaction. We aim to move beyond viewing memory as simple data retention, instead developing conversational agents that build shared understanding with humans, recognize and repair memory misalignments, reflect on their own uncertainties, and help people examine their cognitive frames and biases. In doing so, the dialogue cluster seeks to establish new standards for socially intelligent, ethical, and autonomy-supporting agents, while contributing to real-world domains such as health, cultural literacy, and civic discourse. This ambition distinguishes Hybrid Intelligence from General AI and conventional Human-AI collaboration by emphasizing complementary human-machine intelligence, mutual learning, and sustained reflective interaction.

Internal Collaborations

03

Trust & Trustworthiness (T&T)

Leaders: Myrthe Tielman, Roel Dobbe

Progress in 2025: We have onboarded the 3rd cohort students to the Challenge. We have started discussions around what it means for HI to be trustworthy, and whether this is different from trustworthy AI. As an effort to understand this better, we have started to map the HI projects to the EU ALTAI guidelines for trustworthy AI.

Ambitions in 2026: We aim to use EU ALTAI guidelines and the mapping to it thereof to understand how HI addresses trustworthy AI questions, and potentially contribute this information to the 5-year HI paper or a presentation at a future consortium meeting.

We will maintain our goal of a shared research agenda on trust and trustworthiness in Hybrid Intelligence. Beginning from a set of broad research questions that will be further refined in conversation with the other core members during the project development stage. These questions address what trustworthiness and trust mean across different levels of HI: the technology itself, the interactions between humans and AI, and the wider sociotechnical systems in which HI is embedded. We still aim to understand how trust and trustworthiness can be evaluated both within and across these levels, and how HI systems can be designed to be trustworthy and trusted appropriately. In doing so, this work will lead to a more responsible foundation for Hybrid Intelligence: one that supports effective collaboration without encouraging overtrust, manipulation, or misplaced reliance.

HI Methodologies and Metrics (M&M)

Leaders: Davide Dell'Anna, Tiffany Matej Hrkalovic

Progress in 2025: The MM Challenge started as such in 2025 (before it was referred to as Taskforce). Throughout the year we split the work in six subgroups (called clusters) focusing on specific aspects pertaining to the three main challenge RQs.

Giacomo Zamprogno (PhD in the Challenge) initiated the collection of a dataset of metrics to measure different aspects of HI.

Mireia Yurrita (postdoc in the Challenge) conducted a workshop at the consortium meeting in March 2025 involving 15 members of the consortium to identify their (1) initial ideas and (2) desiderata towards human-LLM agent teaming guidelines (MM Challenge RQ3). This initiative, followed by interviews with other participants, successfully led to a paper published at AAMAS 2026 (also nominated for best paper award) and a dataset of 24 guidelines for the principled design of human-LLM agent teams, adopting a multi-stakeholder approach that includes guidelines for LLM agents, human team members, team designers and embedding organizations.

At the end of 2025, through acquisition of Engineering Funds, we started the development of an online collaborative platform aimed at facilitating the collaborative exploration, refinement, and evaluation of the concept of Hybrid Intelligence (MM Challenge RQs 1 and 2). The platform will primarily serve researchers, including PhD students and postdocs, in the HI consortium, but also professionals and academics working in HI-related fields and the general public, including students, to explore and contribute to what hybrid intelligence is, by reviewing, proposing, and evaluating HI systems and relevant HI metrics.

Ambitions in 2026: In 2026, the challenge will focus on strengthening and connecting its different research lines, using the arrival of a new PhD cohort as an opportunity to foster collaboration, synergy, and a more integrated research community. The core research agenda remains centered on defining Hybrid Intelligence, distinguishing it from related fields such as Human-centered AI, Human-in-the-loop AI, HCI, and traditional AI, developing methods to evaluate the "HI-ness" of human-AI systems, and identifying systematic approaches for designing HI systems. While no external events are currently planned, a possible HHAI 2027 workshop on measuring HI-ness is being considered, pending further internal discussion and PhD involvement. In parallel, the challenge is contributing to the Engineering Fund Project by developing a web-based platform to support research on how HI-ness can be defined and measured.

Special Interest Groups (SIG)

Theory of Mind (ToM)

Leaders: Max van Duijn, Rineke Verbrugge, Ramira van der Meulen

Progress in 2025: We continuously focus on advertising our position within the consortium: Project Intervision, and interesting Reading Groups, in addition to being a platform where people can share their ToM-related work for peer feedback. We would still like to boost our reading group attendance average (around 10 people, currently, three of which organisers) and expand our reach. We have spent the latter half of 2025 brainstorming ideas about less traditional within-consortium outreach (for example, organising a movie night about a ToM-related topic). We offer a platform for our members to share and organise ToM-related events if they wish to do so, but are content with our current schedule of 5 meetings per year (one organisational, three or four reading groups).

Ambitions in 2026: The Special Interest Group for Theory of Mind will continue its yearly activities, including quarterly reading groups in 2026 and an annual in-person meeting during the June/July consortium meeting to reflect on members' research, shared direction, and satisfaction with the SIG's focus. The year will begin with two events: a guest lecture by Reuth Mirsky on disobedience in collaborative AI teams, connected to the ToM4AI workshop community, and a session by Ramira van der Meulen exploring misconceptions about Theory of Mind in AI. The SIG's broader goals remain to map HI projects that use Theory of Mind or related concepts, align background knowledge through readings and discussions, support intervision and consultation on ToM approaches, and facilitate collaboration on concrete ToM projects. With Cohort 3 fully active and Cohort 4 starting, the main ambition for the year is to expand shared ToM knowledge and support new researchers in positioning their work within the field. No major programming projects are currently planned.

NLP4HI

Leaders: Enrico Liscio, Lucia Donatelli

Progress in 2025: We decided to pivot the SIG away from a recurrent meeting event with a fixed pool of participants. Instead, we intend it to be a platform for introducing NLP to the broader HI consortium, through lectures on fundamentals and presentations of HI NLP researchers on their work. The lecture series has been uploaded on the website as educational content.

Ambitions in 2026: The SIG will continue meeting every one to two months, with lectures from HI members who use NLP methods in their research. Enrico and Lucia will coordinate additional activities offline, with the ambition that the group develops toward a shared project by the end of the year. A special session is planned for the HI Consortium Meeting in late June, where the SIG aims to welcome new members of the third cohort and brainstorm possible collaborative activities, such as a joint research paper, hackathon, or another shared initiative. Given the rapid development of NLP, especially with the rise of LLMs, the SIG's main research goal is to help HI researchers stay informed about current methods, best practices, and emerging developments in NLP as a community.

Reinforcement Learning (RL)

Leaders: Shihan Wang, Frans Oliehoek

Progress in 2025: In 2025, SIG-RL welcomed the new third cohort of PhDs and PDs. We continued our regular activities, including internal presentations and community gatherings, also started inviting external speakers. In parallel, we further advanced the development of SHARPIE.

Ambitions in 2026: SIG-RL will continue its monthly meetings with a planned speaker schedule, supported by the HI SIG-RL Google group and Slack channel for announcements, discussion, and community building. A major focus will be the continued development and visibility of SHARPIE, including demo sessions, potential presentation at RLC if the submitted paper is accepted, and plans to develop an HI competition around the platform in collaboration with HI colleagues. The SIG's research agenda centers on the sequential nature of human-AI interaction, emphasizing human behavior simulators, interactive and data-efficient learning techniques, foundation models and domain expertise for bootstrapping, and human-in-the-loop reinforcement learning to support collaboration between humans and RL-based agents. Overall, SIG-RL aims to create synergy by connecting researchers working on related methods and problems, spreading

Ethics and System Design

Leaders: Roel Dobbe

Progress in 2025: The core aim of the SIG is to support cross-consortium interaction and collaboration, particularly with the AlgoSoc and ESDiT consortia. Collaboration with AlgoSoc developed strongly over the past year through exchanges at both executive and researcher levels, including presentations at consortium meetings, shared events, the Destabilize AI Seminar Series, the International AlgoSoc Conference, and the co-located HI–AlgoSoc consortium meetings in March 2026. A key outcome was a shared hackathon on sociotechnical AI challenges, involving early-career researchers from both consortia, which was experienced as highly valuable by both sides.

The collaboration with ESDiT was reactivated in early 2025 with the involvement of postdoctoral

researcher Dr. Caroline Bollen. Together, the SIG organized the “Guidelines for Guidelines” workshop, which explored challenges in developing and using ethical guidelines for AI systems. The workshop focused on interdisciplinary challenges, communication around ambiguities, and constructive ways forward, and may lead to a short position paper using the event as a case study. A further planned collaboration with ESDiT on a longitudinal ethical analysis of the Robotic Surgery use case was explored but has been paused due to limited use-case engagement and time constraints.

Ambitions in 2026: In 2026, the Sociotechnical AI Systems group will continue its biweekly lab meetings to support ongoing exchange and coordination within the research line. Externally, the group plans to organize the Destabilize AI Seminar Series in Fall 2026, with a thematic focus on the climate and ecological impacts of emerging AI technologies, as well as their security implications. The year’s research goals will be further advanced through the recruitment of a new postdoc, who will work on the objectives outlined in Postdoc Project 3.18.

Case Studies

Health – Diabetes

Leaders: Maaike de Boer, Myrthe Tielman, Chenxu Hao

Progress in 2025: The goals for 2025 for the HIT Lab were; This year we will finish a new version of the demo, and we will regroup with the finished PhDs of the first cohort and the start of the third cohort. We will share the demo and the data produced, and we hope that other PhD students will use that as inspiration of their research, which we will accommodate. Looking back, both goals were mostly achieved. We have finished a first version of the demo, and we have had multiple sessions to introduce the use case to the new cohort and make plans for 2026.

Ambitions in 2026: In 2026, the health use case will restart and further consolidate its activities after regrouping around the third cohort of PhD students. A key priority for the coming year is to onboard the new PhD students into the use case and to create a stronger sense of cohesion across the group. The case study team aims to do this through regular group meetings, focused discussions, and small collaborative projects that allow members to connect their individual research interests to the broader use-case agenda. Building on the demo and the data already produced, they will continue developing the use case while also opening up new research directions. For example, the next meeting in January will focus on scenarios and design patterns, helping the group to identify concrete situations, interaction models, and design challenges that can guide future work. The emerging research themes include long-term interaction, reasoning, trust and explainability, and responsibility and stakeholder perspectives.

To create sufficient critical mass and focus, two thematic groups have currently been formed: one around trust and one around long-term interaction. These groups will also incorporate related themes such as reasoning, explainability, responsibility, and stakeholder involvement where relevant. In the first half of the year, both groups will work toward concrete short-term goals, such as developing an architecture, preparing a poster, or creating a demo. Around the summer, they will evaluate the progress and outputs of these groups and decide whether to continue, expand, or reorganize them. Based on this reflection, the case study team will define new concrete plans for the second half of 2026.

Virtual Museum Guide

Leaders: Shenghui Wang, Victor de Boer

Progress in 2025: The collaboration in the case study happens at various levels. At the most generic level, we share common challenges and approaches as well as datasets and other assets. A poster presentation was given at More specifically, with a core group, we have in 2025 been working on a shared challenge and demonstrator: The HI Virtual Museum Guide. This shared work has resulted in a demonstrator presentation at HHAI2025: Ghiotto et al. Supporting the Critical Museum Visitor through Hybrid Intelligence. Next to these shared activities, the individual research projects have contributed to the research challenges. We submitted one Engineering project on Multiplayer VR Data collection. We collaborated with Algosoc researchers to construct a shared HI/Algosoc case around responsible Generative AI in the Museum and Media.

Ambitions in 2026: Our ambition for 2026 is to advance the museum use case into a shared Hybrid Intelligence demonstrator that connects technical development, interaction design, and cultural heritage research. We aim to explore how HI systems can create engaging, ethical, and inclusive museum experiences through facial mimicry, polyvocal decision support, responsible elicitation of user perspectives, and intelligent robotic interaction in exhibitions. Through monthly meetings, hackathons, collaboration with relevant matrix elements, a possible international workshop on “HI for Museums,” and an engineering project proposal, we seek to strengthen the use case as a concrete platform for studying Hybrid Intelligence in cultural and public-facing settings.

Robotic Surgery

Leaders: Chirag Raman, Dalibor Vasilic, Catholijn Jonker

Progress in 2025: This year we made some major progress towards getting the Twin OR Capture Lab to functional status. We are in the process of collecting the first dataset using the facilities as part of two MSc thesis projects at TU Delft. We are in the process of writing the publication about the capture lab and preparing the code and tools for public release. We also had many 3rd cohort Ph.D. students start with the case study this year. During the case study meetings, we gave all the students a tour of the Twin OR

facilities along with a tech demo, described the case study goals and brainstormed with them about how they can validate ideas relevant to their projects within the case study. Additionally, the DiMiRoS (Digital Microscopy Redesigned) project data set is being reviewed and prepared for publication in NeurIPS. This is a collaboration between Erasmus MC (Dalibor, Cadey Tang, and team) and TU Delft (Chirag, Anuj Singh; a 3rd cohort PhD student).

We also had fruitful discussions with the Museum Case Study (Shenghui) about aligning pipelines for digital human representations within our case studies. This is something we will be investigating further within the Engineering Grants round in 2026.

Ambitions in 2026: In 2026, the TwinOR case study aims to move from infrastructure development toward active investigation of robotic assistance in surgical settings. Through monthly meetings, open demonstrations at the TU Delft Capture Lab and Erasmus MC operating room settings, and continued work on realistic surgical scenarios, the case study will explore how robotic assistants can understand surgeons' verbal, gestural, and nonverbal behavior, and whether these capabilities can generalize across surgeons. A second ambition is to assess whether the current multimodal sensory setup is sufficient for situational awareness in the operating theatre, and to identify what additional sensing may be needed. Alongside ongoing publications on the Capture Lab, DiMiRoS data, the case study position paper, and the first Capture Lab dataset, the team will propose an extension of the TwinOR engineering grants. This next phase will focus on using simulation for agent perception and planning, improving the human reconstruction pipeline in alignment with the Museum Case Study, and integrating a BCI interface to support new third-cohort projects.

Scientific Assistant

Leaders: Pradeep K. Murukannaiah, Ilaria Tiddi

Progress in 2025 The work by students in the use case led to a paper on Automatic Evaluation Metrics for Artificially Generated Scientific Research. The paper also produced a resource by parsing all OpenReview submissions into a unifying format, augmenting them with additional metadata. This year, we started a collaboration with an industrial partner – Elsevier. The goal is to study the scientific assistant problems from the stakeholders' perspective and propose Hybrid Intelligence methods to solve them. We currently have 8 defined 3-months miniprojects ("spikes"), two of which have already started, and the others to be starting between Q3-Q4 2026 and Q1-Q2 2027. Accordingly, we defined a few joint subchallenges: Perform a multi-modal, explainable collaboration. Operate on data with different modalities at different scales. Extract (multimodal) content from text. Maintain research integrity (information quality): how to build an HI-collaborative team to detect fraudulent/fake papers. How to improve scientific indicators (provenance challenge). How to maintain Trust and trustworthiness, i.e. Actor/action-based trust vs. artefact-based trust.

Ambitions in 2026: For 2026, the research aims to strengthen Hybrid Intelligence in the scientific process by developing multimodal, explainable, and trustworthy methods for scientific collaboration. Key goals include generating tailored and interactive explanations, supporting reviewer assignment, extracting relevant context from text and metadata, evaluating revised hypotheses, and improving information quality, provenance, and scientific indicators. The work will also further develop living meta-reviews, improve their coverage and generalizability, and contribute to a more formal understanding of "HI-ness" in science through neuro-symbolic architectures, a Hybrid Intelligence ontology and thesaurus, and methods for maintaining trust and trustworthiness.

Law

Leaders: Cor Steging, Stephanie Tan

Progress in 2025: As the Hybrid Law case study started in late 2024, the year 2025 has been primarily focused on creating a clear and sustainable direction. Some of the efforts of 2025 have already come to fruition at the time of writing (early 2026). We explored various stakeholders and opportunities, and have broadened the understanding of the AI & Law field within the Hybrid Law group through meetings and talks, to answer our first research question (RQ1). These ideas and visions have been written down as Blue Sky paper in early 2026. Towards the end of 2025, we have focused on close collaboration with the National PoliceLab, and we have set up a roadmap to develop a fraud-intake system that can aid ordinary citizens that may potentially be affected by fraud. In a co-design session with police officers, we have worked on a prototype ruleset. This trajectory aims to answer research question 2 (RQ2). At the same time, we have and still are actively exploring connections with other legal entities, such as the National Forensics Institute, in order to see how Hybrid Intelligence can assist legal experts (RQ3). Roadmap was created to design the fraud intake system. Initial prototype rulesets have been created during the co-design session.

Ambitions in 2026: For the upcoming year, the Hybrid Law research line aims to consolidate and expand its work on Hybrid Intelligence for legal and justice-related contexts. Through monthly law leads meetings, bi-monthly group meetings, and focused writing sessions, the group will advance both its conceptual agenda and concrete research outputs, including an HHAI BlueSky paper and the SATIKAS paper. A central research ambition is to reproduce and extend the fraud analysis system by incorporating neurosymbolic methods and improving user interactions, while also reflecting on the opportunities and limitations of collaboration with the Police Lab and the National Police. The group will further explore additional legal use cases and strengthen external engagement through invited speakers, PoliceLab meetings, collaboration with the AlgoSoc Justice sector hackathon, participation in the Dutch AI-in-law community, and a national workshop on Hybrid Law toward the end of the year. Programmer-related work will focus on the HI Law engineering project proposal and the SATIKAS project.

Governance

Organization-wise, Catholijn Jonker assumed the role of coordinator on 1 January 2025 together with a 0.2 fte full professorship on HI at VU, with Frank van Harmelen becoming vice-coordinator. Erman Acar and Nazmiye Ozkan joined the Management Team.

The year was defined by the **launch and recruitment for the 3rd research cohort**, with projects well-distributed across the matrix elements of Collaborative Synergy, Dialogue, and Trust. The matrix underwent a few changes, notably, seeing the addition of the HI Law case study.

Collaborative efforts with other major programs, specifically **ALGOSOC**, were formalised through joint workshops and the planning of a combined Hackathon for 2026. On the technical front, the consortium completed its **second round of engineering funding**, supporting projects like the “Bubble Board” as a collaborative platform, while launching a **joint HI GitHub repository** to centralise shared code and data.

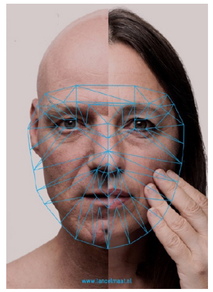
The three consortium meetings were well visited by around 80 participants each. Furthermore, the consortium held a pivotal strategy meeting focused on **institutionalising the “Science of HI”** and preparing initiatives for sustaining the Hybrid Intelligence Centre beyond the end of the grant and improving our societal impact.

The Executive Board **released internal funds to hire 7 postdocs** to ensuring research continuity. We saw promotions for Frans Oliehoek (to full professor at TUD), Pradeep Murukannaiah (to associate professor at (TUD), and Catha Oertel (to associate professor at TUD).

Lastly, Leon Eshuijs represented his fellow PhD’s in the Executive Board meetings, and carried out a survey to investigate what the PhDs thoughts were on the HI consortium, including what parts they liked most and what could be improved.

Building the community

Artists in Residence program was launched, with Kim Baraka leading the initiative to foster the dialogue between scientists and artists. **Karen Lancel and Herman Maat** have been selected as the artists in residence.



The consortium finalised its advisory structures by completing the Scientific Advisory Board and installing a Societal Advisory Board featuring representatives from government (Amsterdam and Rotterdam), industry (KPN, ASML, Exin, Improven), consulting (Deloitte, KPMG) and healthcare (Universiteit Maastricht).

HI-Collective was established by Wendy Aartsen and Jos Mitsunaga in June 2025, directly inspired by the HI consortium to introduce the Hybrid Intelligence approach to companies.

Staff arriving

New staff from participating organisations have joined the Hybrid Intelligence Centre:

1. Suzan Verberne

We welcome them to the HI consortium as the new supervisors of the next cohort PhD and Postdocs.

Writing retreats

Finally, 2025 was the second year in which our Program Manager Wendy Aartsen hosted several writing workshops for small groups in secluded circumstances in a cottage in nature. These writing groups were very fruitful and contained several training elements, such as one-one conversations with the host and peer feedback. These were especially useful for last-year PhD students working on their PhD thesis and for postdocs and assistant professors working on grant proposals

The Organisation 04

Education & Training

05

Course development and PhD course teaching

Our PhD students follow courses in the Netherlands Research School SIKS (School for Information and Knowledge Systems) and, to a much smaller extent, the Kurt Lewin Institute (Social Psychology). There is a close collaboration with SIKS: Most PhD students are members and follow several SIKS courses a year. In 2025, several courses developed by HI and SIKS together were taught, some of them on site, others online.

The HI + SIKS course 'Explainable AI' (XAI) was adapted from the 2021 online version by course directors Anske Fokkens and Rineke Verbrugge and taught a second time, this time on site in Vught, on 2-3 June, 2025. Part of the lecturers were the same as in 2022 and part were new. This time, the course ended with a very successful poster session were PhD students presented their own XAI-related research.

In 2025, a new bi-yearly HI + SIKS course has been developed by an HI senior, Charlotte Gerritsen, and a colleague at the VU, Michel Klein, on 'AI and Human Behaviour'. AI & Human Behaviour is a prominent area of research concerning AI systems that support individual people in voluntarily adapting their behaviour. Such technological support can be relevant in different domains like health, sustainability and justice, e.g., to help people adopt healthier lifestyle patterns, make more sustainable choices, or manage their chronic disease. In this two-day course, the lecturers investigated key components of AI-based behaviour support systems. Key to such technologies is understanding why people do what they do (by learning about their motivations, habits, capabilities and needs), so that the offered support is timely and targeted at a pivotal mechanism. AI-related technologies can contribute to deepening this understanding, e.g., machine learning of observational data to gain insights in behavioural patterns, cognitive models to reason about cognitive aspects such as motivation and self-efficacy, and conversational AI such as chatbots to engage with people. The HI + SIKS course 'Reinforcement Learning for Adaptive Hybrid Intelligence', of which the first instalment had taken place in 2023, was adapted by the course directors Herke van Hoof, Erman Acar and Shihan Wang and taught for the new generation of PhD students on site in Vught on October 20-21, 2025. The course included exercises to allow participants to integrate reinforcement learning in their own research.

Additionally, the HI + SIKS course 'Responsible AI', first taught online in 2022, was adapted to current developments by course directors Virginia Dignum and Rineke Verbrugge and taught online. The students appreciated the course, especially the hands-on exercises in which they had to role-play stakeholders in a high-stake decision on the use of HI and to analyze one another's PhD projects for responsibility aspects and possible solutions to ethical risks.

Preparations were started in 2025 for the development of several of the HI + SIKS courses that were to be taught in 2026, among others, a new course on Knowledge Graphs and a course on 'The Science of HI.

PhD progress

PhD-students (as well as postdocs) prepared a brief annual report at the end of 2025, in a standard format which does not only request reporting on the research questions, progress and results, but also on the coherence with other projects, so as to trigger the students to look beyond their own projects. Extracts from these reports are included as the final chapter in this report. As is customary at Dutch universities, all PhD students who have been employed for a full year have had a progress-advice meeting with their first (internal) and second (external) supervisors; every PhD student in the Hybrid Intelligence Center has a second supervisor from another participating university.

PhD reading group

Bi-weekly PhD meetings are organised, alternatingly online and on-site. The goal of those meetings is to create a common conceptual basis as well as a social community among PhD students from diverse fields ranging from computer science through social psychology to philosophy, and to train them in leading group discussions about HI-literature. As work progressed, PhD students also gave presentations about their research. Students receive training credits (represented on annual personal certificates) for participation, leading a reading group session, presenting their research, and co-organizing events.

In 2025, the HI PHD reading group meetings again saw a large change in membership: several PhD students of the third cohort joined the group and introduced themselves and their research.

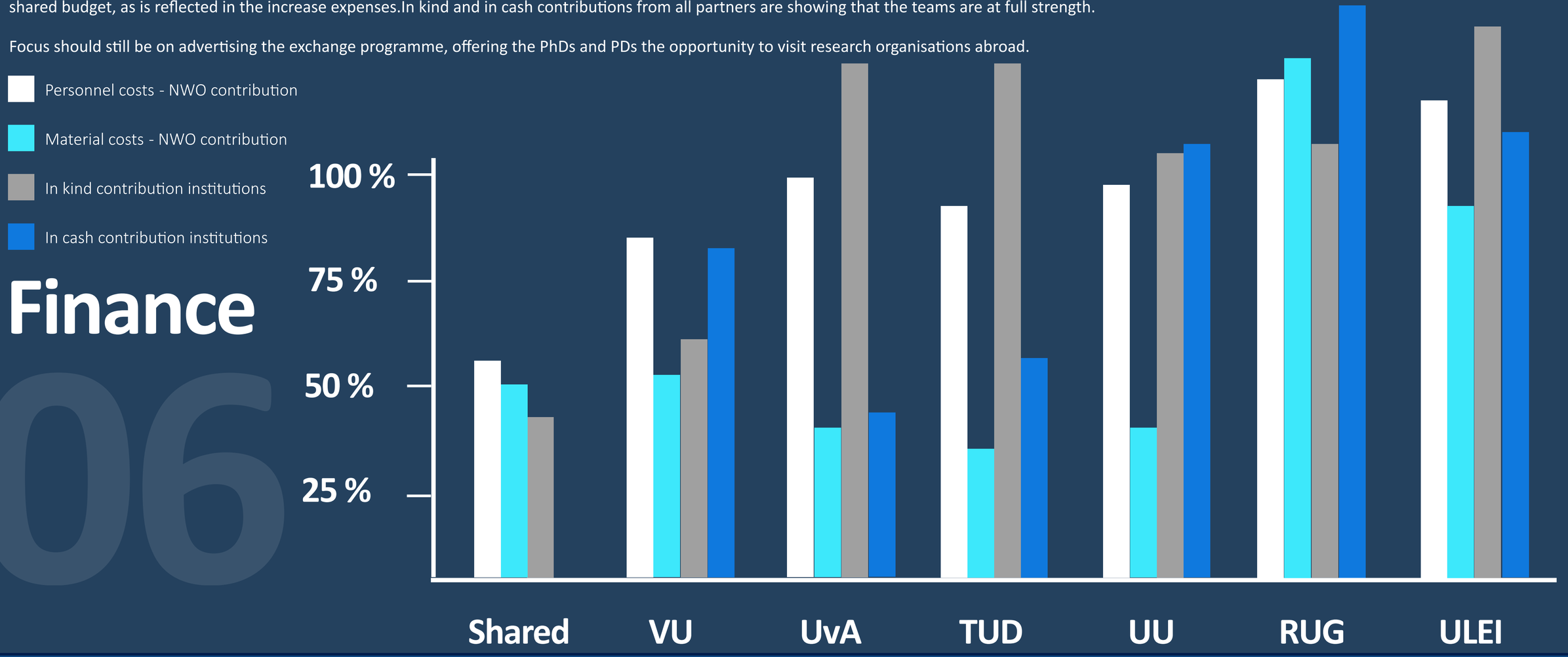
Realisation Hybrid Intelligence 2025

Within the Hybrid Intelligence Centre we started the recruitment for the 3rd and 4th cohort of PhD and PDs in 2025. The finances reflex that most partners have onboarded new team members or are in the process of doing so. This year the awared engineering projects were executed at the engineering teams of the VU and the TUD. As the software being develped is serving the consortium Case Studies their costs are booked at the shared budget, as is reflected in the increase expenses.In kind and in cash contributions from all partners are showing that the teams are at full strength.

Focus should still be on advertising the exchange programme, offering the PhDs and PDs the opportunity to visit research organisations abroad.

Finance

06



Michiel van de Meer on BNR Nieuwsradio: Voting advice from language models

Michiel van de Meer appeared live on BNR Nieuwsradio to discuss a timely and critical topic: the role of language models in providing voting advice for the Dutch general election.

Roel Dobbe Publishes Critical Opinion Piece on the Global AI Arms Race in NRC: he argues that the global AI arms race is creating a modern version of the natural resource curse—one that fuels environmental destruction, deepens inequality, and undermines human rights.

Catholijn Jonker on Hybrid Intelligence at SIGNAL ’25: At SIGNAL ’25, a tech-focused event hosted at the H’ART Museum, Catholijn Jonker gave a lecture titled Building the Bridge Between Humans and AI. Speaking to an audience of CTOs, engineering leads, and product and IT directors, her talk focused on how Hybrid Intelligence can support human decision-making in software development and programming.

Working Visit of the Dutch Parliamentary Committee on Health to TU Delft

On November 25, the Parliamentary Committee on Health, Welfare and Sport (VWS) visited TU Delft to discuss the challenges and opportunities in the healthcare sector. The key message: the healthcare system is under significant pressure and requires action—but technological innovation alone is not enough.

Frank van Harmelen on the Knowledge Graph Insights Podcast

Frank van Harmelen joins Larry Swanson for episode 33 of the Knowledge Graph Insights podcast. In a wide-ranging conversation, Van Harmelen discusses the promise of hybrid human-machine intelligence and the emerging role of neuro-symbolic AI.

VU Highlights Hybrid Intelligence Centre, In an interview, professor Frank van Harmelen discusses how collaboration between humans and AI is becoming increasingly important, and how the Hybrid Intelligence Centre is putting this theme firmly on the scientific agenda.

Interview with NEMO Kennislink: Frank van Harmelen discusses the transformative potential of Hybrid Intelligence, emphasizing how the collaboration between humans and AI can lead to major scientific breakthroughs rather than replacing human input.

Frank van Harmelen Keynote at TNO’s Appl.AI Splash Event 2025: On April 22 in Utrecht, Frank van Harmelen delivered a keynote as Principal Investigator at TNO’s Appl.AI Splash Event 2025.

JUST LAUNCHED : HI PODCAST

Hybrid Society has a provocative podcast exploring the uneasy co-evolution of humans and artificial intelligence.

Hosted by Joshiya Mitsunaga, with co-hosts Prof. Catholijn Jonker (Professor of Interactive Intelligence at TU Delft) and Prof. Frank van Harmelen (Full Professor of Knowledge Representation & Reasoning, Vrije Universiteit Amsterdam)

External Outreach

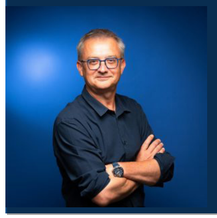
07

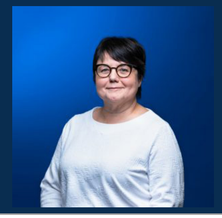
08 International conference

250 participants
5 days
13 workshops
5 tutorials
38 main conference papers

The Fourth International Conference on Hybrid Human-Artificial Intelligence in Pisa, Italy organised by the National Council of Italy and Scuola Normale Superiore and University of Pisa.

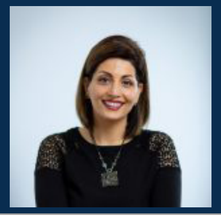
HHAI2026 is organised by Université Libre de Bruxelles, Vrije Universiteit Brussel and AI For the Common Good Institute Brussels in Brussels, Belgium, as the fifth conference on Hybrid Human Artificial Intelligence.


Tom Lenaerts


Ann Nowé

Shenghui Wang

André Meyer-Vitali,


Maryam Alimardani

Joost Vennekens

Anca Stefano Florea

Eladio Montero

Tim Brys

Ambitions Year 6

09

Alumni

With the 1st cohort finishing their projects and taking on new roles (some within HI, others outside of HI), and similarly some of our academic staff taking on new roles outside HI, the urgency for setting up an alumni programme is growing. This year, we aim to reach out to our alumni and set up with them a number of activities that connect them to each other and to new HI activities.

Scientific ambition

With the third cohort hired and active, we now are ready to bring in and onboard the 4th cohort. The matrix leads and executive board members played an essential role in connecting the scientific goals of the 3rd and 4th cohort project proposals with the goals of our Case Studies, Special Interest Groups and Challenge groups (our “matrix elements”) and to the overall goals and missions of the HI consortium. Through a dedicated onboarding process for the 3rd and 4th cohort researchers within these matrix elements, we aim to strengthen the coherence between the matrix elements and the individual projects. The ambition of the matrix leaders, the executive board and the management team is to have the supervisors, postdocs and PhDs work together to **reach the third level of the HI capability table**, and to collaboratively focus on **show-casing the added value of HI.**

The Next Generation Programme

A major ambition for year 7 is to launch a Next Generation programme. **The Next Generation programme** is a three-layered initiative, consisting of the academic layer, the applied layer, and the political layer. Within the academic layer we aim to spawn new research projects, and to disseminate the HI mission by educational activities. The grant writing will support both the fundamental and the applied sciences of HI, while advancing the careers of the next generation of HI researchers. In these activities the senior researchers will advise and encourage the younger HI researchers to take the lead as coordinators of these proposals. The educational ambitions consist of developing postdoctoral HI courses, introducing master level courses on H within existing AI and computer science programmes, and ultimately establishing dedicated HI master programmes. In the political layer, the goal is to bring the HI philosophy to national and EU-level governmental organisations with the ambition to influence AI policies on both the national and European levels.

10

PhD/PD projects

Summaries

Project 1.01, Aishwarya Suresh

AI, Moral Testimony and Deference

Goal: The project aims to develop an account of the limitations of AI systems in supporting humans in forming moral judgements.

Research Questions in 2025: I explored three interrelated topics for my PhD, of which I was able to resolve two of them in 2025.

1.Moral Artificial Enhancement: I looked at the state-of-the-art discussion of Moral Enhancement with AI. I was hoping to develop a better way of assessing AI's ability to help humans make moral judgements. And I was able to do that by arguing for treating AI as demonstrating a kind of expertise, which can be delivered to people in the form of testimony.

2.Psychological Implausibility of Pure Moral Deference: I looked at the state-of-the-art discussion on moral deference. I was hoping to make sense of this narrow and near-implausible concept of pure moral deference. And I was able to do that by building an argument to reject the psychological plausibility of pure moral deference.

Progress in 2025: I was able to finish the drafts of two of my papers in 2025, and I am close to finishing the draft of my third and final paper in a week. I am hoping to finish and submit my PhD by September 2026

Project 1.06, Emre Erdogan

Computational Theory of Mind for Collaborative Privacy

Goal: The long-term goal of this research is to design and develop a formal, feasible, and human-inspired computational Theory of Mind (ToM) for effective Hybrid Intelligence**.**

Research Questions in 2025:

1. How can AI agents embed ToM reasoning with sensitivity to values (e.g., privacy) to resolve conflicts and foster group alignment in multiagent settings?

2. How can AI agents govern the relations between different types of social dynamics that are essential to collaboration, and use them with ToM reasoning to adapt their interactions with humans?

Progress in 2025:

1. We developed and evaluated a computational ToM agent architecture that incorporates value-sensitive reasoning based on insights from an empirical user study. Through multiagent simulations, we showed that ToM-enabled agents, especially those with learning capabilities, significantly improve group alignment and reduce multi-party privacy conflicts compared to baseline approaches.

2. We designed a formal framework for modeling how social dynamics such as trust and respect dynamically evolve and influence one another in collaborative settings. This framework equips agents with relational operators and reasoning rules that, combined with ToM, enable them to interpret and adapt to human social dynamics over time. We demonstrated its applicability through a privacy-sensitive content sharing case study, showing how agents can better align their interactions with human expectations**.**

Project 1.07, Kevin Godin Dubois

Combining NeuroEvolution and Reinforcement Learning for Embodied Robots

Goal: The objective is to combine evolutionary approaches with online learning using the modular robots developed by the VU lab.

Research Question in 2025:

What are the benefits of human–in-the-loop reinforcement learning on agents' accuracy when trained in deceptive environments?

What are the important dimensions when considering perception of modular robots by the general public?

Are Multi-Layer Perceptrons necessary for accurate control of modular robots?

Progress in 2025: The first question was uncertainly answered with curriculum training performing about as well as human-guided one. The second is shaping up to be a human study performed alongside B. Hilpert. Third one points toward very shallow MLPs being preferable to larger ones, in terms of performance and efficiency. However, physical replication is, at this stage, not feasible due to dropped/unavailable hardware support for the previous/new platform.

Project 1.08, Niklas Höpner

Hybrid approaches to learn explainable sequential behaviours

Dissertation Title: Algorithms for knowledge-guided sequential decision-making

Abstract: Reinforcement learning and imitation learning have become foundational frameworks for learning control, particularly within a two-stage pipeline combining large-scale pretraining with reward-based fine-tuning. This thesis investigates mechanisms to enhance the scalability and generalization of this pipeline across embodied, commonsense, and scientific AI domains. In the embodied AI setting, we address the pretraining data bottleneck in two ways: first, by developing a semi-supervised model that extracts task segments from unstructured videos, achieving performance comparable to five times the labeled data; and second, by introducing a cross-agent framework using a shared diffusion planner to pool data from diverse embodiments. In the fine-tuning stage, where knowledge is typically transferred via pretrained weights, we instead propose transferring knowledge by integrating knowledge graphs directly into deep reinforcement learning algorithms. This approach exploits object class hierarchies to compose policies at multiple levels of abstraction, substantially improving generalization to unseen objects. Finally, for scientific agents, we explore the potential for defining a reward signal for fine-tuning. By curating a large-scale dataset of metadata-annotated peer reviews, we demonstrate that specialized scientific embedding-based models predict citation and review scores more reliably than large language models, suggesting their potential utility as a reward signal for fine-tuning.

Project 1.09, Bram Renting

AutoAI for dynamic data

Dissertation Title: Learning in Automated Negotiation

Abstract: This dissertation advances automated negotiation by developing agents that can learn and adapt across diverse negotiation settings through three increasingly sophisticated approaches: automated algorithm configuration, portfolio-based strategy selection, and end-to-end reinforcement learning with graph neural networks. We demonstrate that while each of our developed methods successfully pushes the boundaries of what negotiation agents can achieve, our end-to-end reinforcement learning approach shows particular promise in reducing human-induced bias while maintaining conceptual simplicity. In the second part, we evaluate negotiating agents through extensive empirical research, including organising the Automated Negotiating Agents Competition (ANAC) and demonstrate that learning agents generally outperform non-learning agents. Our analysis reveals fundamental limitations in standard evaluation metrics for negotiation agents, particularly showing that rankings based on average utility are highly dependent on opponent group composition. The dissertation concludes by proposing multi-agent meeting scheduling as a concrete application domain that could provide clear performance criteria and drive meaningful progress in automated negotiation research.

Project 1.10, Anna Kuzina

Continual Learning and Deep Generative Modelling for Adaptive Systems

Dissertation Title: Mapping the Invisible: Insights into Deep Latent Variable Generative Models

Abstract: This thesis contributes to the field of deep latent variable generative models. We provide a comprehensive background on deep generative modeling, covering different model types, neural network parameterization, and learning algorithms. Advancements in latent variable models are proposed by enhancing both model density estimation capabilities and the quality of the learned representations. Based on these contributions, the thesis is structured into two parts.

The first part addresses research questions related to enhancing density estimation performance by exploring better probabilistic modeling approaches. First, we focus on a trainable prior distribution and an optimal prior defined by the prior works. Based on this concept, we develop a training framework for continual learning of the variational autoencoders and propose a deep hierarchical VAE which exhibits superior performance and better training stability. We next analyze diffusion models, another class of latent variable models. We study their generative and denoising abilities and propose DAED, a combination of the diffusion model and denoising autoencoder, where these two functions are explicitly decoupled.

The second part explores properties of the learned latent representations. First, we show how the hidden data representation learned by the model can be made robust to adversarial attack. Second, we explore the ability of the latent variable model to preserve symmetries of the data and show how these can benefit downstream applications.

Project 1.12, Taewoon Kim

A Machine With Human-Like Memory Systems

Goal: This project addresses the phenomena of identity, reference and perspective within communicative scenarios between AI and people in real-world situations. It aims to build personal relationships between people and AI in an adaptive environment to learn how to communicate about shared situations or contexts.

Research Questions in 2025:

-How can AI maintain a persistent, interpretable memory over time?

-What representations (e.g., temporal knowledge graphs) best capture evolving context?

-How can symbolic memory be combined with learning-based methods?

-What are the trade-offs between neural and structured memory systems?

Progress in 2025: The work in 2025 resulted in multiple research manuscripts presenting the proposed approach and initial experimental findings. These were submitted to peer-reviewed venues. Although not accepted, the feedback has been used to improve the clarity, evaluation, and framing of the work.

Project 1.13, Annet Onnes

Monitoring AI systems using Bayesian Networks

Goal: Main: How to use Bayesian Networks for normative monitoring of a Black-Box AI system?

-How to use Bayesian Networks to measure a Black-Box AI system's deviations from acceptable behaviour?

-What is an effect threshold when measuring a Black-Box AI system's deviations from acceptable behaviour using Bayesian Networks?

-How to capture stakeholders' norms and expectations in a Bayesian Network?

Research Questions in 2025:

How to quantify Bayesian Networks using maximum entropy and constraints elicited from stakeholders?

How to formalise the modulation reasoning pattern as an idiom for Bayesian Network construction?

Progress in 2025: My work in 2025 firstly focussed on the research for a particular paper working on the question: How to quantify Bayesian Networks using maximum entropy and constraints elicited from stakeholders? This paper was submitted and accepted for presentation at the 18th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU) in September. Additionally, I worked on an extension to a previous project (accepted and presented at a conference in 2024), formalising a reasoning pattern as an idiom for Bayesian Network construction, using my extended definition of idioms. This work was included in a thesis chapter. Following the input received after the presentation at ECSQARU, I revisited and extended the quantification of BNs using entropy. Here the result was also an update and extension to a thesis chapter.

Project 1.14, Pei-Yu Chen

Interactive Machine Reasoning for Responsible HI

Dissertation Title: Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support

Abstract: This thesis investigates how conversational interaction, specifically through alignment dialogue, can help behavior support systems better understand and adapt to a user's real-time needs. Behavior support systems are often deployed to help individuals manage health-related goals, such as increasing physical activity or adhering to dietary guidelines. However, these systems tend to rely on pre-collected behavioral data, which makes it difficult to adapt support to the nuanced, changing circumstances of daily lives. We thus introduce alignment dialogues as conversations between an AI system and a user aimed at eliciting reasons for non-adherence, surfacing underlying values, and revealing contextual barriers. Rather than merely logging behavior, these dialogues help uncover the user's lived experience to iteratively build a more accurate user model at run time.

After conducting an exploratory focus group study, we found that probing for deeper knowledge like user values often triggers negative emotional responses, suggesting that the interpersonal impact of an agent's language is a critical design dimension. These findings informed a large-scale user experiment testing different dialogue strategies, which highlighted a tension between conversational depth and usability: while open-ended questions allow for greater perceived completeness, they also increase cognitive effort compared to multiple-choice formats. To assess how these outputs could be used in healthcare, we explored sharing dialogue content with professionals as a Conversational Intermediary AI (CIAI). Interviews and experiments suggest that while these agents facilitate early diagnoses and sensitive discussions, the effectiveness of the information is sensitive to the presentation format and length of the dialogue.

Finally, we demonstrate technical feasibility through a system prototype that transforms user responses into structured knowledge using RDF triples and a user knowledge graph. This allows the system to cross-reference user data with medical expertise to recommend actions that are contextually relevant and clinically sound. Our findings show that the effectiveness of alignment is determined not only by the accuracy of the information gathered but by the interpersonal and emotional quality of the interaction. This approach redefines personalization from categorization to interpretation, where systems must learn to make sense of the user's situation. Ultimately, this research suggests that personalization grounded in dialogue opens new paths for hybrid collaboration, advocating for behavior support systems that adapt by conversing, reflecting, and aligning with users.

Project 1.15, Urja Khurana

Predicting generalizability of NLP experiments

Dissertation Title: From Concept to Impact: Toward More Robust Language Model Deployment

Abstract: The rapid adoption of large language models in daily life can have a significant impact on society. For especially safety-critical applications, e.g., hate speech detection, misclassifications can be disastrous; calling for robust and reliable behavior in real-world conditions. Such applications also demand a shift from the conventional machine learning workflow for language models: data selection, model development, and evaluation. In most cases, the workflow does not reveal actual model behavior but mere general insights into performance through an aggregated score. Each stage of this workflow requires a deeper analysis to ensure that application-specific criteria are met when a model is deployed. In this thesis, I conduct such an end-to-end analysis and go on a journey through the machine learning workflow, with hate speech detection as a primary use case. In Part I / Chapter 2, I first aim to get a thorough understanding of the task of hate speech detection to guide data selection, which is influenced by deployment context. Through insights from social science and law, I propose a framework for hate speech practitioners that encapsulates the aspects essential for hate speech and can be used to create deployment-based hate speech definitions and annotation guidelines. In Part II, I revisit whether current model setups are suitable for tasks that go beyond a single notion of the ground truth. While it makes sense for a model to make a prediction when there is clear human consensus, it does not for cases with high subjectivity. In Chapter 3 I propose a subjectivity-informed method that allows a model to abstain when it estimates low crowd agreement. The results show that this method is especially beneficial for unseen data. In Part III, I discuss model evaluation; understanding its generalization capabilities. A robust model not only performs well on data it has seen during training but also on data in real-world conditions. This is especially essential for a task like hate speech detection. The phenomenon of underspecification can hamper this demand: models trained under the same circumstances and achieving similar scores can still show diverging behavior. Therefore, I focus on stability; exhibiting the same capabilities across different seeds. In Chapter 4 I start by measuring stability through model agreement on mistakes and aim to improve stability by using a cheap ensembling method that averages model weights across the gradient descent trajectory. As the technique does not appear to improve much in stability or performance, in Chapter 5 I take a data-centric perspective to investigate when this particular ensembling method could yield improved performance and stability. The hypotheses built around its effectiveness under different data properties are not supported by the results. As a next step, I conduct a preliminary study in Chapter 6 on the relationship between stability and the certainty of a model. The initial results on the task of question answering suggest a positive correlation between stability and certainty depending on the question type for each model, prompting a thorough investigation for the future. We then explicitly tie the three main components of the workflow together. It is important to verify whether the intentions (intended type of hate speech) are preserved throughout the workflow and is characterized by the model that will be deployed. In Chapter 7 I propose a framework to investigate whether the behavior of a model aligns with the intended definition of the hate speech dataset on which it was trained. When applied to hate speech datasets, the framework reveals gaps between the intended task definition and model behavior. The methodologies proposed across the workflow are a step toward more robust and reliable deployment of language models.

Project 1.16, Kata Naszadi

Context modelling for dialogue systems

Goal: In this project, we will investigate the role of context (agent or environment) on natural dialogue generation. We will explore several research directions such as the detection and representation of knowledge gaps, the generation of contextually appropriate responses, and novel ways to represent and access large contextual information.

Research Questions in 2025:

Are humans and large language models aware of when an instruction asks for something explicitly as opposed to when certain underspecifications are resolved using context?

Do agents (humans or artificial) adapt their interpretations based on how reliable pragmatic inferences are with a certain partner?

Progress in 2025: We were finishing a human data collection for a block-world building experiment that tests if humans adapt to pragmatic failures of their partner by starting to lower their confidence in pragmatic interpretations. The manuscript was submitted and is under review for Cognitive science. We confirmed strong adaptation affect from only task success as feedback. The preprint is available on psyarxiv. The data is available and published . We extended this experiment by also testing large language models on the same task, with the addition that we implemented an agent that can answer questions from the LLMs, this way the LLMs could express uncertainty in the form of clarification questions instead of certainty ratings. This task is designed to test if LLMs are aware of what's being explicitly said and what is being inferred from context. The preprint is available on arxiv and is under review.

Project 1.19, Wijnand van Woerkom

Explaining data-driven decisions with legal, ethical or social impact to end-users

Dissertation Title: A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law

Abstract: Case-based reasoning (CR) involves comparing new problems with previous cases to support decision-making. In common law systems, this forms the core of the stare decisis principle: judges and lawyers rely on precedents to justify their reasoning and arguments. An influential model of CR is the result model (RM), which formulates a principle of a fortiori reasoning: if an earlier case was decided in favor of a party, then any new case that is at least equally favorable to that party must also be decided in its favor. This model and its variants are foundational to this dissertation. An analogy with artificial intelligence (AI) is central here: just as judges generalize from precedents, AI systems generalize from training data. This makes it possible to use the RM and its variants to analyze data-driven AI systems. Part I investigates how the RM can be extended into a general theory of a fortiori CR, incorporating hierarchies, dimensional values, and incomplete information. This results in variants such as the hierarchical RM and the dimensional-hierarchical RM. Part II builds on these extensions and focuses on applications of the theory in AI and law. The theory has been implemented in logical software and tested on several datasets. One case study concerns the COMPAS system, a controversial tool used to predict recidivism. The analysis demonstrates how the theory can contribute to transparency and the responsible use of AI.

Project 1.20, Merle Reimann

Enabling co-regulation for long-term engaging semi-structured conversations

Dissertation Title: Speaking the Same Language: Spoken capability communication in human-agent and human-robot interaction

Abstract: When people interact with a specific conversational agent or robot for the first time, the capabilities of that system are not necessarily clear to them. To enable even non-expert users to have smooth and efficient interactions, they have to be informed about the conversational agents / robots capabilities. One way of doing this is through the interaction itself. In this thesis, I focus on spoken interaction. I contribute to this by investigating how conversational agents' and robots' communication of their capabilities influences the user's mental model and behavior. For this, I improved a conversational agent with a level-based evaluation method, which allowed for targeted improvements. I then used this agent to investigate how the users' mental model change over time and depending on whether or not the agent communicates its capabilities. In contrast to conversational agents, robots are still an unfamiliar tool for many. Because of this, it is important to be able to identify whether a novelty effect makes users interact out of curiosity or if they are interacting because of an actual need. I propose indicators derived from a study in the wild that can help with this identification. The study showed that the novelty effect is still a common reason for interactions, which indicates that people do not yet have correct knowledge about the robot and its capabilities. Therefore, I compared two different capability communication strategies with a baseline and found that people did change their interaction style if they were proactively informed about the robot's capabilities. With the work presented in this thesis, I am contributing to making conversational agents and robots understandable by human users, enabling them to have smoother and more efficient interactions.

Project 1.21, J.D. Top

Recognition of non-cooperative behavior

Goal: In hybrid intelligent systems in which humans and agents interact together, it is important to be able to understand and detect non-cooperative behavior such as lying and other forms of deception, both in people and in software agents. Cognitive scientists have shown that understanding the concept of lying and being able to maintain a lie over time requires second-order theory of mind: reasoning about what the other person (or agent) thinks you think. The same level of theory of mind is also required for detecting whether others are lying to you. In this PhD project we will:

- investigate the logical and computational foundations of deception and deception detection in hybrid groups;
- lay the theoretical groundwork for modelling and analyzing non-cooperative behavior in several communicative contexts such as negotiation games and coalition formation games;
- develop principled methods for the design of software agents that can detect when other group members are engaged in non-cooperative behavior such as lying;
- build agent-based models and/or computational cognitive models of deception and deception detection;
- use simulation experiments in order to predict the outcomes of lab experiments to be performed.

The objective of the temporary position is the production of a number of research articles in peer-reviewed scientific journals and conference proceedings, which together will form the basis of a thesis leading to a PhD degree (dr.) at the University of Groningen.

Research Questions in 2025: Can we develop a task that can be used to effectively study cooperation and deception in the context of theory of mind? Within such a task, how often do participants use deception, and at which stage of play do they deceive? Which theory of mind orders do participants use? Which norm is established for cooperative play? Can we develop a computational agent that is not distinguishable from human play?

Progress in 2025: The paper ``Cooperation, Deception and Theory of Mind in a Cyclic Game with Inter-Player Signalling.'' introduces the novel Mod-Signal game, a two-player, 24-action, non-zero-sum variant of rock-paper-scissors where the numbers 1 through 24 are laid out on a circle. One player first signals their to-be-played action. Signals are non-binding. Then, both players simultaneously select an action. If your number is one higher than the other player's, you gain a point, where 1 is one higher than 24. In our behavioural experiment we find that while cooperation is faster and yields more points, participants still play primarily deceptively, using no more than second-order theory of mind when doing so.

In Cheryl's Birthday puzzle, Cheryl gives a list of ten dates: May 17, May 18, June 14, July 16, July 18, August 15, August 16, August 17, September 14, and September 15. Cheryl gives Albert only the month, and Bernard only the day. Everyone is a truthful and perfect reasoner, and all of this is common knowledge. Albert then announces ``I don't know when Cheryl's birthday is'', after which Bernard announces ``I know when Cheryl's birthday is.'' When is Cheryl's birthday? The paper ``How Well Do People Perform on Novel Logic Puzzles Requiring Higher-Order Theory of Mind?'' introduces variations of this puzzle that require first- through fourth-order theory of mind to solve. Experimental results show that participants are slower and more error-prone for higher-order puzzles, performing worse than chance level at puzzles that require fourth-order theory of mind.

Publications:

Top, J.D., de Weerd, H., Raveenthiran, A., Jonker, C. M, & Verbrugge, R. (2025). Cooperation, Deception and Theory of Mind in a Cyclic Game with Inter-Player Signalling. Proceedings of the Annual Meeting of the Cognitive Science Society, 47. Retrieved from https://escholarship.org/uc/item/1466g95n

Minculescu, A., Top, J.D., Verbrugge, R., & de Weerd, H. (2025). How Well Do People Perform on Novel Logic Puzzles Requiring Higher-Order Theory of Mind? Proceedings of the Annual Meeting of the Cognitive Science Society, 47. Retrieved from https://escholarship.org/uc/item/5s63j43q

Project 1.22, Íñigo Martínez de Rituerto de Troya

Including ethical and legal considerations in the HI design process

Goal: The project focuses on understanding how human factors are crucial for informing the design of AI systems in making sure these are safe in their behavior, in their treatment of people, and allow affected stakeholders to be empowered to hold the system and its managers/operators/developers accountable.

Research Questions in 2025:

RQ1. How is AI safety negotiated and realised in public policy AI systems?

RQ2. How can a sociotechnical systems view explicate the shortcomings of current safety efforts?

Progress in 2025: Preparing dissertation for submission to committee.

Project 1.23, Andreas Sauter

Causal Reinforcement Learning and Reasoning

Goal: In this project, we will design an adaptive agent which will enable causal inference through learning and reasoning in a reinforcement learning environment. In particular, the agent is expected to combine ontological background knowledge and causal discovery and inference effectively in achieving a given task in a partially observable environment. We will try to obtain theoretical guarantees, and outline its performance empirically.

Research Questions in 2025: Can we achieve principled, amortized causal inference via generative models? Can such results be applied to time series data?

Progress in 2025: We designed a generative model architecture to query causal questions in a zero-shot manner. We also started experiments on applying this method to financial time series data.

Project 1.24, Maria Heuss

Explainable De-biasing in Learning from Interactions

Dissertation Title: Responsible Advice Giving Systems- Fairness and Interpretability in Information Retrieval

Abstract: As AI systems increasingly provide automated advice across hiring, healthcare, and conversational applications, they raise critical questions about responsible deployment. This thesis addresses two fundamental challenges: ensuring fairness across individuals and groups, and enhancing explainability to make systems more interpretable for developers and users.

In the first part of this thesis we examine the fairness in ranking systems by challenging two fundamental assumptions. First, it addresses the misconception that user exposure patterns are always predictable, developing a method to reduce unpredictable exposure distributions while maintaining system utility. Second, it tackles the assumption of perfectly accurate relevance estimation. We demonstrate how model uncertainty can help to balance user utility with fairness objectives through reordering documents based on uncertainty levels, thereby significantly reducing bias in top-ranked results with minimal utility loss.

In the second part we examine interpretability in advice-giving systems through two approaches. First, we extend feature attribution to ranking systems, introducing RankingSHAP for more flexible analysis of ranking decisions with competitive performance. Second, it investigates citation faithfulness in retrieval-augmented generation (RAG) systems, distinguishing between faithful citations that actually influence answers versus merely correct ones containing relevant information. Empirical analysis reveals widespread unfaithful citations and problematic post-rationalization behaviors in state-of-the-art RAG models.

Our work highlights the critical need for thorough evaluation of proposed approaches and careful examination of prevailing assumptions in current practices. While we have made meaningful progress on individual components, achieving truly trustworthy and reliable advice-giving systems will likely require a more holistic perspective that integrates fairness, interpretability, and other dimensions of responsible AI deployment.

Project 1.25, Sharvaree Vadgama

Teaching artificial neural networks how to reason

Goal: The aim of this project is to model, study, and visualize decision-making processes in (convolutional) neural networks. In particular, emphasis is put on explainability in critical applications where reliability and trustworthiness of AI systems are crucial, such as in the medical domain. The main theme will be the design of AI systems through recurrent NNs where focus is put on modeling (and visualization of) the dynamics of decision making rather than relying on the static black box mechanics of feed-forward networks.

Research Questions in 2025:

- How do equivariant architectures compare to scaling large non-equivariant methods?
- Understanding applications of equivaraint architecture in different applications

Progress in 2025:

- **Literature survey on Geometric Generative frameworks** (yet to be published). Compiled a comprehensive review of geometric generative models and their applications, covering the landscape of symmetry-aware generative modeling frameworks and their use across scientific domains.
- **Clifford algebraic structure in generative models** [2] This work applies Clifford Group Equivariant Neural Networks to generative diffusion models for 3D molecular generation. Rather than relying solely on scalar and vector representations, the models operate on the richer multivector representations of the Clifford algebra, which are capable of capturing more expressive geometric structure inherent in molecular systems.
- **Scaling equivariant transformer architectures across tasks** [1] — Two complementary papers address this question. This work introduces the Platonic Transformer, which defines attention relative to reference frames from Platonic solid symmetry groups, inducing principled weight-sharing that enables combined equivariance to continuous translations and discrete Platonic symmetries while preserving the exact architecture and computational cost of a standard Transformer.
- Submitted two workshop proposals accepted at ICLR 2026 (submitted in 2025)

Project 1.26, Loan Ho

Knowledge Representation Formalisms for Hybrid Intelligence

Goal: Knowledge graphs (KGs) have a significant role in storing and providing access to global knowledge, common and accessible to both human and artificial agents, representing knowledge for different agents in Hybrid Intelligence (HI) settings. Unfortunately, providing suitable formalisms to work with complex, conflicting, dynamic and contextualised knowledge is still a big challenge. This project aims to propose an approach to represent and adapt knowledge dynamically in different scenarios (user preference, available sources, dynamic or contextualized knowledge).

Research Questions in 2025:

R1: How to address inconsistencies in Knowledge Graph?

R2: How feasible is a proposed framework that addresses inconsistencies in large-scale Knowledge Graphs?

R3: How to evaluate a proposed framework that addresses inconsistencies in real-world application?

Progress in 2025:

To answer R1, I proposed a novel structured argumentation framework to compute and explain logical reasoning results in (inconsistent) knowledge base. The result of this was submitted to the Argument&Compute journal.

To answer R2, I conducted a set of experiment to evaluate the feasibility of the proposed framework in large-scale knowledge graphs (DBpedia, LOV, Wikidata, LUMB university). The result of this was submitted to ISWC 2026

To answer R3, I conducted a user study to evaluate the effectiveness of the proposed framework (e.g., satisfaction and user understanding). The result of this was submitted to HHAI 2026.

Project 1.28, Atefeh Keshavarzi Zafarghandi

Game-Theoretic Development and Analysis for Hybrid Intelligence

Goal: To develop a pipeline for structural argument mining that extracts proposition units from text and classifies them as claims or premises using multi-agent LLM reasoning, with a focus on improving role classification through persuasion-based dialogue game rather than fine-tuning.

Research Questions in 2025:

1. Can multi-agent LLM interaction improve proposition-level argument component classification compared with single-pass or discussion-only prompting?

2. Does a persuasion-style structural challenge--proof protocol outperform a standard multi-agent discussion framework for distinguishing claims from premises?

3. How much does proposition segmentation quality affect downstream argument component classification?

4. Can structurally grounded, inference-time prompting provide a useful alternative to fine-tuning in small-data argument mining settings

Progress in 2025: We built a multi-stage argument mining pipeline for short argumentative texts. The pipeline first segments raw text into proposition-like units and then classifies each proposition as a claim or a premise. We implemented and compared two multi-agent LLM role-labeling approaches: a baseline discussion framework and a persuasion-based framework in which labels must be defended through explicit structural links. We evaluated the system on the English Arg-Microtexts corpus using XML-derived gold labels and alignment-aware evaluation. The persuasion-based approach improved role-labeling performance over the discussion baseline, especially on structurally ambiguous examples involving objections, counter-considerations, and pro/con argumentative moves.

Project 1.31, Mark Adamik

Commonsense reasoning for embodied agents

Goal: This PhD project will tackle the problem of enabling intelligent systems to reason over common sense knowledge to achieve their tasks. We will focus on the specific case of embodied agents performing everyday activities such as cooking or cleaning, i.e. tasks that require reasoning over different types of knowledge (we call these knowledge-intensive tasks), and using for this purpose the available common sense knowledge graphs, i.e. structured data sources that contain knowledge about the everyday world. We will propose new methods to allow robots to achieve knowledge-intensive tasks through combining existing structured data (knowledge graphs) with situational personalised knowledge captured through their sensors. We will study the hybrid application of advanced symbolic and subsymbolic AI methods to robotic systems operating in complex real-world scenarios.

Research Questions and Progress in 2025: The research conducted in 2025 focused on how semantic knowledge representations can support the automatic generation of symbolic planning structures from perceptual and linguistic inputs. In particular, the work investigated how ontologies and knowledge graphs can be used to bridge the gap between perception and classical planning by enabling the automatic construction of PDDL planning artifacts. The core research question underlying this work can be summarized as follows: how can structured semantic knowledge and perceptual observations be combined to automatically generate planning domain and problem representations for robotic systems? This question addresses a key bottleneck in robotic task planning, namely the need for manually engineered domain files and hand-crafted problem descriptions. The work explored whether ontologies could provide the formal structure required to reconstruct planning models, while visual-language models and perception modules provide the contextual information necessary to instantiate concrete planning problems.

Project 1.32, Selene Baez Santamaria

Common-sense knowledge representation & reasoning for embodied agents

Dissertation Title: Knowledge-centered conversational agents with a drive to learn

Abstract: We propose a framework for adaptive knowledge-centered conversational agents (KCCAs) that acquire knowledge from conversations, evaluate its quality, and are internally driven to become more knowledgeable. Unlike service agents trained for predefined tasks and based on unidirectional interactions, KCCAs treat people as diverse, social sources of information across open domains. Each KCCA is equipped with an episodic Knowledge Graph (eKG): a structured, evolving memory that organizes knowledge communicated in dialogue into RDF triples and nested named graphs. This information is accumulated in a database that stores claims, related perspectives, and original sources along with contextual and temporal information about the dialogues. This representation enables KCCAs to manage subjective knowledge, detect conflicting viewpoints, and reason about people's beliefs over time. The eKG also supports domain-independent assessment of knowledge and interaction quality. Structural patterns in the graph can be associated with knowledge dimensions such as correctness, completeness, redundancy, interconnectedness, and diversity. Applying these patterns to a KCCA's memory results in targeted areas for quality improvement that can be regarded as the agent's knowledge needs. Moreover, tracking the evolution of the eKG provides insight into knowledge exchange during dialogue and correlates with human judgments of interaction quality. KCCAs are also equipped with task-agnostic dialogue capabilities for knowledge acquisition, such as dialogue management, natural language generation, and natural language understanding. To actively pursue its knowledge needs, KCCAs employ reinforcement learning to adapt their communicative strategies. To express their current knowledge, they aggregate information in a perspective-aware manner, allowing for nuance where needed. To ingest new knowledge, they transform natural language into RDF triples to be added to the eKG. We demonstrate the KCCA framework in three real-world applications, including: 1) Type 2 diabetes lifestyle support, 2) Personal timeline reconstruction, and 3) Counter-speech generation. These use cases highlight the strengths of graph technologies for dialogue agents, particularly due to their structure and connectivity properties. We further extend the KCCA framework to include embodied agents, developing tools for capturing and creating multimodal interactions. Together, these contributions provide a dynamic, adaptive, and socially aware foundation for Hybrid Intelligence systems capable of long-term, knowledge-driven interaction.

Project 1.33, Lea Krause

Contextualised Natural Language Generation within the framework of collaborative agents

Dissertation Title: Contextualising Conversational AI

Abstract: Conversational AI has evolved from simple rule-based systems to sophisticated large language models capable of engaging in complex dialogues. However, despite significant advances in fluency and coherence, these systems often struggle with fundamental aspects of human communication, particularly in how they handle context as a pragmatic factor. This thesis investigates how conversational AI systems can better understand, represent, and respond to the multiple layers of context that shape human dialogue. Through an examination of uncertainty management, perspective integration, and context representation, we explore both the potential and limitations of current approaches to context-aware conversational AI. We begin with a theoretical foundation grounded in Grice's Cooperative Principle, establishing how pragmatic maxims influence current AI implementations. The first research direction explores how conversational AI models handle uncertainty, examining both the expression and consistency of uncertainty across different contexts. Through multilingual experiments and analysis of explanation triggers, we find significant misalignments between models' internal confidence levels and their expressed certainty, particularly

in non-English languages where models maintain high confidence despite degraded performance. The second direction investigates how systems incorporate individual perspectives, examining the integration of subjective information, user-specific adaptation mechanisms, and cultural considerations. Research into task-oriented dialogue systems reveals challenges in balancing factual and subjective knowledge, while analysis of multilingual datasets shows that current approaches often preserve Anglo-centric biases rather than representing diverse cultural viewpoints. The final direction examines how conversational AI can be grounded in world and personal contexts through structured representation mechanisms. We explore Graph-based approaches for maintaining dialogue memory and context across interactions, demonstrating promise for both transparency and long-term coherence in conversational systems. The findings show that while conversational AI has made substantial progress, current approaches often fail to capture the nuanced ways humans use context to shape communication, from expressing appropriate uncertainty to adapting responses based on cultural backgrounds. We conclude that context should be treated not as an optional enhancement but as a fundamental component of conversational AI design, requiring research that bridges theoretical understanding of human communication with practical implementation in AI systems.

Project 1.35, Masha Tsfasman

Memory for Conversational Agents

Dissertation Title: Towards predicting memory in multimodal group interactions

Abstract: People often remember parts of conversations that are important to them, such as something personal, useful, or emotionally engaging. These memories help shape relationships, guide decisions, and influence how we communicate in the future. While many computer systems can already track emotions or attention in group settings, no previous research has looked at how specific moments in conversations are stored in memory or how this process could be predicted using technology.

On the path towards training computer systems to predict such memorable moments, this dissertation first introduces a new dataset called the MeMo corpus (Chapter 2). It includes group video conversations along with direct reports from participants about which moments they remembered. The data was collected in a way that reflects real-life conversations, using repeated video calls and memory reports that are linked to specific moments in time.

The study in chapter 3 then asks whether affective signals, such as emotional tone or energy in a conversation, could help predict what people will remember. These kinds of emotional signals are often used in artificial intelligence systems. However, the results show that emotional signals alone are not enough to explain what people remember from a conversation.

Next, in chapter 4, the dissertation looks at other behavioural signs, such as where people were looking and who was speaking. These signals were found to be significantly linked with memory: for example, people tend to remember parts of a conversation where there was shared attention or dynamic speaking patterns. Using these signals, simple computer models were able to predict which parts of the conversation were more likely to be remembered. The study also looked at why people remembered certain moments and found that many of them were related to personal relevance or social connection.

This work shows that it is possible to build systems that recognise which parts of a conversation are more memorable. This can be useful for improving automatic meeting tools, personal assistants, and other technologies that support communication and augment memory.

Project 1.36, Morita Tarvirdians

Developing a computational model of perspectives for human-robot interaction

Goals: Conversational AI has the potential to support individuals in their thinking processes, specifically in cognitively demanding situations such as high-stakes personal decisions. However, despite recent advancements in conversational AI, it still lacks computational models of how individuals think and reason, which hinders the development of AI tools that augment human reasoning rather than replacing it. A critical aspect of modelling this is understanding how individuals structure and express their thoughts, and which aspects of personal context they draw upon when reasoning. This project focuses on capturing and quantifying individual thought patterns in pre-decision contexts to build agents that can model and adaptively support human reasoning through human-agent dialog

Research Questions in 2025: Main RQ:

Can a conversational agent adaptively support integrative pre-decisional reflection by personalizing its dialog strategy to individual users' thought profiles?

The overarching goal was to design and evaluate a novel user modelling approach embedded in a conversational agent, and to develop adaptive dialog strategies that build upon this model to support integrative unique reasoning trajectories rather than steering users toward solutions or homogenizing reasoning across individuals.

Progress in 2025: We introduced ReflectiMate, a conversational agent that builds a real-time user model of individual thought patterns and applies an exploration-exploitation algorithm to adaptively broaden and deepen reflection. ReflectiMate was designed and implemented as a standalone system, which is available as a private GitHub repository and can be made available to other researchers upon request.

To answer the research questions, a between-subjects study (N=128) was conducted on Prolific with human participants.

This work has been accepted at UMAP 2026 (User Modeling, Adaptation and Personalization), an A-ranked conference in the field.

Data from the study will be made available in accordance with the Data Management Plan and HREC agreements

Project 1.43, Deborah van Sinttruije

Conversational AI for Social Good

Goal: This research targets at the development of an argumentative dialogue system for human-robot interaction, which will explore how to fuse verbal and non-verbal behaviour to infer a person's perspective. I use, and further develop, reinforcement learning techniques in order to drive the robot's argumentative strategy for deliberating topics of current social importance such as global warming or vaccination. We build a POMDP model for an agent that adjusts to human working memory capacity over time in a cognitive task. We tested the POMDP model in a user study.

Research Questions in 2025: The research focused on how AI systems can better support human performance by adapting to people's cognitive abilities in real time. The research questions we tried to answer are: how can a Partially Observable Markov Decision Process be used to model and adapt to these hidden states over time; to what extent do such adaptive agents outperform traditional performance-based heuristics; and how dynamic factors such as stress and fatigue can be incorporated into these models. To address these questions, the work develops and compares adaptive strategies, and validates these approaches through user studies, with the goal of designing AI systems that maintain an optimal level of challenge, prevent cognitive overload, and enhance human performance in a principled and transparent way.

Progress in 2025: A controlled experiment, a dataset capturing behavioral and subjective measures, and a paper that is currently in preparation for journal submission.

Project 1.46, Sergey Troshin

Controllable probabilistic models of language

Goal: The main goal of this project is to develop probabilistic generative models with application to natural language processing. As the first step, we focus on modeling uncertainty in representational space and introduce a new geometry-aware probabilistic model designed for language representations (embeddings). Next, we focus on the decoding from large language models (LLMs), targeting specifically LLMs as probabilistic models. Main practical focus includes constrained language generation, which means satisfying external constraints, a direction most aligned with the HI research directions, namely Responsible HI (trust via uncertainty estimation, constrained generation); other practical applications include ambiguity resolution; LLMs and creativity, prevention of mode collapse; study of quality-diversity trade-offs, and quality-efficiency trade-offs for test-time compute.

Research Questions in 2025:

How to control the exploration-exploitation trade-offs for LLMs during test-time?

How to improve the sampling efficiency of LLMs by switching to non-i.i.d. Sampling?

Progress in 2025: We propose a test-time decoding adapter, which helps us to use the model's internal features to decide whether to prioritize exploration (high temperature) or exploitation (low temperature) during each step of text generation.

We propose and formulate a new paradigm of sequential (non-i.i.d.) sampling, which allows us to treat classic parallel sampling and sequential processing using a common framework, and demonstrate the higher diversity (sampling efficiency) of sequential sampling techniques. This work is easily applicable and extendable, and inspires me to continue this direction further (ongoing projects).

Control the Temperature: Selective Sampling for Diverse and High-Quality {LLM} Outputs, Asking a Language Model for Diverse Responses

Project 2.01, Dylan Jia-Huei Ju

Assistant for exploring scientific literature

Goal: We seek to support the investigation step of the scientific method cycle through a scientific assistant that helps identify relevant literature (at the document level) and pinpoint relevant facts within it. Beyond finding relevant documents, the assistant targets the coverage of relevant facts (i.e., a fine-grained unit of information) across collections. To this end, we aim to investigate retrieval methods that ensure retrieved results collectively address user information needs, rather than returning topically similar documents that convey redundant facts. The assistant will learn from user feedback to refine its representation of information needs and better locate relevant facts within the returned literature.

Research Questions in 2025:

- How can we measure retrieval quality systematically with human perspectives for scientific search assistants ?
- How does retrieval quality affect long-form outputs generated by LLM ?
- What retrieval strategies best improve information coverage for long-form generative assistants ?

Progress in 2025: We investigated how retrieval quality affects long-form retrieval-augmented generation (Long-form RAG) outputs and how to measure and improve information coverage for long-form RAG. We introduced an evaluation framework using human-written summaries and question-based evaluation to assess retrieval coverage beyond standard relevance metrics. To actively improve coverage, we proposed an LLM-based reranker that uses sub-question prediction to maximize nugget coverage. We further propose a dense retrieval method trained with coverage-aware objectives that achieves better nugget coverage over strong retrieval baseline methods.

Project 2.02, Floris den Hengst

A RL-based assistant for Patient-centric Decision Making

Goal: This project aims to develop training methods for an RL agent to learn a strategy that is (i) cooperative with the expert, that (ii) displays behaviour which is coherent from the human point of view, and that (iii) takes into account the often conflicting goals that a medical treatment is expected to satisfy. We want to develop novel techniques in the RL framework based on the medical requirements of learning a treatment strategy and consequently demonstrate this by a prototype.

Research Questions in 2025:

1. How to incorporate multiple stakeholders preferences into sequential decision-making?
2. How to construct human-understandable expressions that explain dependencies in a dataset at different locations of the data distribution?
3. To develop a platform to support controlled experiments into hybrid human-AI sequential decision-making involving human participants
4. How can AI models advise humans in debiasing their language?

Progress in 2025:

1. Formalized an initial model for incorporating multiple stakeholders preferences in sequential decision-making under linearity assumptions.
2. Developed and empirically validated an approach for regression with closed-form symbolic models (formula) at user-specified quantile levels.
3. Lead an engineering initiative to develop a platform to support controlled experiments into hybrid human-AI sequential decision-making involving human participants.
4. Developed a dataset and approach for the detection of linguistic bias in government documents.

Project 2.03, Leon Eshuijs

Specializing NLP models through reinforcement learning

Goal: In this project, we explore the possibilities of using RL to target specific behaviors of models. This could be training specifically to avoid a social undesirable bias, but also to exhibit robust behavior on a specific phenomenon (e.g., negation, specific syntactic structures, temporal information interacting with truth values)

Research questions in 2025: The research questions I'm trying to answer for 2025:

1. How do LLMs come to their decision when finetuned under data with spurious correlations and can we reliably trace this decision process to improve their faithfulness?

2. When LLMs are optimized via RL they've shown to lead to undesirable behavior such as reward hacking and targeted manipulation, but how does this RL training step get influenced by the safety pretraining the model under goes?

Progress in 2025: The main results is a publication at the conference CONLL, which was collocated with ACL in 2025. I presented there my paper a mechanistic interpretability investigation of shortcut learning in LLMs.

A second summary is that a large part of my most recent paper was concluded in 2025, a paper which is currently under submission and given the positive reviews and with some luck will be published later this year. In this work, which was also partly concluded in 2026, we investigated factors for harmful misalignment due to RL optimization of LLMs, and found that the model's safety training can act as a safety buffer limiting the harm from faulty RL optimization. (paper available on arxiv here: https://arxiv.org/abs/2604.12500)

Project 2.05, Bernhard Hilpert

Affective signals in interactive reinforcement learning

Goal: In this project we investigate if and how robot affective expression, grounded in RL, can help to make the learning process more transparent. Previous research has shown that robot emotion simulation grounded in RL is possible and simulation studies show that the resulting emotional intensities (joy/distress/hope/worry) are plausible from a psychological perspective.

Research Questions in 2025: In 2025, I am trying to answer the following RQs:

RQ 1: In Human Interactive Robot Learning (HIRL), what are the mental models that human teachers construct when observing Reinforcement Learning-powered learning robots?

RQ 2: In Human Interactive Reinforcement Learning, what are the intuitive teaching approaches that human teachers take when observing RL-powered learning robots?

RQ3: Based on RQs 1 and 2, what are the misalignments in teacher mental models of robot learners that create suboptimal teaching approaches?

RQ4: What are the affective signals that could bridge these misalignments identified in RQ4?

Progress in 2025: In 2025 I published the first half of results of my qualitative experiment in which I open-endedly prompted participants to report on their intuitive mental models at the AAMAS 2025 conference and the co-located ALA workshop.

Further, I analyzed the results of RQ2 from the same experiment, which resulted in a teaching framework for HIRL settings (submitted) and identified 3 misalignments from the joint results of RQ 1 and 2.

Lastly, I built a modular framework to simulate and evaluate affective signals based on these misalignments.

Project 2.06, Gülce Günaydin

Semantic technology for aggregating research studies

Goal: Cooperation Databank (CoDa) is the first databank to represent an entire field of studies in the social sciences in a machine-readable format that allows users to conduct on-demand meta analysis. Our goal is to create living meta-analytical reviews of various topics covered by CoDa to present a comprehensive summary and most up-to-date statistical results in the cooperation literature.

Research Questions in 2025: Our goal was to further develop living reviews produced by CoDa by expanding the range of topics it covers and improving its structure. We aimed to improve the templates that would allow us to produce multiple new living reviews efficiently.

Progress in 2025: We worked on defining the queries in a way that would allow us to produce living reviews efficiently. This effort resulted in the development of templates that allow us to generate multiple living reviews.

Project 2.07, Feline Lindeboom

Algorithmic support for online deliberation

Goal: We will develop a methodology for supporting participatory democratic decision making online. The methodology should support (1) users in participating in online democratic deliberation, (2) policy makers in aggregating users' choices and opinions in a principled and transparent manner.

Research Questions in 2025: Can we develop and study (axiomatically, algorithmically and maybe also empirically) sensible group based diversity notions for approval based committee elections? Such theory can help make diverse summaries of deliberations, highlighting different viewpoints.

Progress in 2025: I had my first paper accepted at AAMAS 2026 (see below). It is titled “Diverse committees with incomplete or inaccurate approval ballots”. In 2025, I have been working on multiple other projects of which 2 have survived so far. One is on diversity axioms in approval based committee elections (a subarea of multiwinner voting, part of computational social choice theory). One is on detecting clones in portionment (also voting theory, computational social choice). Both have applications in deliberation summarization. The first has been accepted at the Social Choice and Welfare Conference 2026, but this is a conference without proceedings so there is no publication (it was a first version, the project is still running).

Project 2.08, Alejandro García Castellanos

Collaborative annotation of training data for deep learning

Goal: We will develop techniques for collaborative human-computer image annotation of training sets for deep learning tasks. Our techniques will suggest relevant annotations to the human annotator, will indicate inconsistencies in the human annotations, and will use concepts from geometric deep learning to handle shapes of image annotations.

Research Questions in 2025:

1. How to incorporate geometric information on neural approaches from geodesic segmentation?

- Neural Eikonal Equations can enable collaborative annotation systems via geodesic segmentation. Current approaches use simple PINNs and Neural Operators. What are the benefits of framing neural Eikonal solvers as conditional neural fields?

2. How can we obtain the maximum expressivity of geometric hierarchy discovery methods in hyperbolic space?

- Hyperbolic space has been shown to be a natural choice for hierarchical data. We want to test how to maximize its benefits when constructing hierarchies via Steiner Minimal Trees. These hierarchies can be used for phylogenetics, but also for some computer vision tasks such as hierarchical segmentation.

Progress in 2025:

Equivariant Eikonal Neural Networks: Grid-Free, Scalable Travel-Time Prediction on Homogeneous Spaces [Accepted at NeurIPS 2025]

- Here we tackle RQ1. We presented a grid-free Eikonal solver for arbitrary Riemannian manifolds that can quickly compute solutions and encode them compactly, providing significant utility for seismic travel-time simulation and weak segmentation via geodesic segmentation.

Randomized HyperSteiner: A Stochastic Delaunay Triangulation Heuristic for the Hyperbolic Steiner Minimal Tree [Accepted at AISTATS 2026]

- Here we tackle RQ2. We showcase how by tacking an stochastic heuristic we can improve the performance of deterministic heuristics for constructing SMTs, by up to a a 32% improvement.

Project 2.09, Mustafa Mert Çelikok

Bayesian multi-agent reinforcement learning for learning to collaborate with humans

Goal:.This project aims to investigate how Bayesian methods can be used in multi-agent reinforcement learning (MARL) as a way for an AI agent to infer models of its human partners, in order to collaborate with them more effectively and with as few interactions as possible.

Research Questions in 2025:

1. How can we build scalable experimentation platforms to systematically evaluate interactive, bidirectional communication patterns between multiple reinforcement learning agents and human participants?

2. Under what conditions can an AI agent learn sample-efficient, zero-shot cooperative meta-strategies by observing a population of socially rational, self-interested partners with private utilities?

Progress in 2025:

1. Framework Development: Co-developed SHARPIE (Shared Human-AI Reinforcement Learning Platform for Interactive Experiments), a versatile, Python-based web framework encapsulating Gymnasium APIs to standardize and log multimodal human-agent collaboration.

2. Game-Theoretic Complexity Foundations: Formulated the concept of “socially intelligent populations” (incorporating consistency and pairwise compatibility). Provided formal lower bounds showing zero-shot coordination is impossible without prior population data, and introduced an “imitate-then-commit” meta-strategy establishing stronger sample complexity upper bounds than standard imitation learning.

Project 2.10, Chenxu Hao

Automatically identifying decision-making moments during social interactions for better partner selection and collaboration

Goal: The goal of this project is to bridge the artificial perception and action loop by considering the idea of building a HI Social Guru. This is carried out through (i) goal setting, (ii) enlightenment and (iii) empowerment via the HI Social Guru. The important thing about a Guru is that they assist humans in self-reflection in social settings without needing to interfere directly (synchronously) with the interaction. That is, they provide wisdom and enlightenment asynchronously from the interaction. This provides another take on how HI systems can help humans to perform better in social situations.

Research Questions in 2025:

1. How do users' actual goals and behaviors diverge from the assumed task goals in social interactions?

2. What behavioral, conversational, or reflective indicators can reveal task incongruence in social interactions?

3. How can task-aware reflection methods support deeper learning in contexts such as cooperation tasks and medical communication training?

Progress in 2025: In 2025, the research examined task incongruence in social interactions for technology-supported reflection. In one project, we studied partner selection in a cooperative task and showed that users sometimes made decisions that diverged from the assumed task goal, even in structured and incentivized settings. We also developed a LLM pipeline for annotating conversational data and explored whether LLM-based analysis of conversation patterns could indicate such incongruence. In another project, we extended this idea to Breaking Bad News medical communication training. We identified a mismatch between communication goals, checklist-based reflection prompts, and what learners actually reflect on. Together, the projects established Task-Incongruence Awareness as a foundation for designing reflection systems that support deeper understanding of users' goals, behaviors, and learning needs.

Project 2.11, Giacomo Zamprogno

Hybrid argumentation using large scale knowledge graphs

Goal: Develop a way to use large-scale structured data harvested from the (structured, machine-readable) web in argumentation about complex hypotheses.

Research Questions in 2025: How to use argumentation-inspired methods and structured knowledge to make knowledge discovery methods more robust.

Progress in 2025: Survey paper on Scientific Assistant tools and how they fit within the HI dimensions.

Research paper on rule-based link prediction extended with semantic constraints.

Collaboration on a paper for the HI Scientific Assistant use-case.

Project 2.12, Saloni Singh

Fostering synergetic exchange for creative human-AI collaboration

Goal: How can an AI system foster synergetic exchange with a human during a creative process? We will develop a collection of novel algorithmic tools that will allow a human+AI system to achieve an improved outcome when collaborating together as opposed to when engaging in a creative process by themselves.

Research Questions in 2025:

Literature Review Research

How are human-AI co-creative systems designed, both in terms of interaction patterns and underlying AI methods? What common themes and design approaches emerge across existing co-creative systems? How do these design choices influence the creative process, collaboration experience, and user engagement?

Experimental Study Research

How do creative outcomes differ between AI-assisted and human-only narrative planning conditions? How does the framing of AI suggestion sentences (e.g., neutral vs. curiosity-inducing) influence user engagement during narrative planning tasks?

Progress in 2025: Literature Review Paper (Accepted with minor changes), a first experiment paper written and submitted, and a conference visit for a poster.

Project 2.13, Ramira van der Meulen

Modelling Common Ground for Theory of Mind

Goal: The aim is to study how common ground emerges, and how such common ground can support Theory of Mind, using agent-based modeling of a collaborative game and lab experiments in which humans and AI agents play that same game.

Research Questions in 2025: Overarching research goals of 2025:

- How is the formation of Common Ground affected by the use of different solution strategies (within the same environment)?
- In what way do different solution strategies affect the collaboration between individuals? Why would one solution strategy be chosen over the other?
- What is the status and application of Theory of Mind within the field of AI?
- Can Theory of Mind be formalised into a broadly but universally usable concept for AI research based on past literature from different fields?

Progress in 2025: The largest direct output in 2025 has been the 'Four Misconceptions about Theory of Mind: Reviewing the Use of Theory of Mind in Artificial Intelligence and Computing Science' paper, in which we discuss the status of Theory of Mind in Artificial Intelligence, through a lens of a literature analysis in Development Psychology, Evolution literature, Linguistics, Neuroscience, Philosophy, etc., to give the field of AI a strong literature backing of topics to take into account when they wish to relate the concept of ToM to an AI project.

Additional efforts have been made towards the Social Models of Intelligence (SMoI) and the Hybrid Vertical Learning projects, both explorations of the way the human brain can learn, absorb and execute strategies for problem-solving. SMoI focuses on the different affordances, failure modes, costs/benefits of various solution strategies (we compare 'learning by trade-off' (mindless partner accommodation vs game theoretical optima), 'Theory of Mind' and 'norm-following behaviour' as solution strategies. In Hybrid Vertical Learning, a major study for 2026, we analyse whether solution strategies can be non-verbally transferred from generation to generation (i.e., teacher-learner, learner becomes the teacher of the next generation).

Project 2.14, Lena Malnatsky

Hybrid content generation for personalized child-robot interaction

Goal: To develop and study a hybrid approach for designing personalized educational responsible child-robot interaction aimed at longer term. Before LLMs, creating child-appropriate robot interaction content was largely a manual craft process: researchers and designers handwrote the robot's personality, conversational goals, narratives, dialog structures, and many of the utterances needed for individual sessions. With the rise of LLMs, an opposite risk has appeared: many approaches now over-automate this work, which can reduce pedagogical control and weaken human oversight over child-facing content.

My research aims to balance these two extremes. I investigate hybrid solutions tailored to educational needs, using AI when it is genuinely useful, while keeping humans in the loop for both authoring and oversight. This hybrid logic is applied throughout the process: we seek co-designed solutions by involving children, educators, and parents in design and evaluation; we use LLMs carefully and selectively; and we explore more responsible forms of CRI design, for example by making robot memory a co-constructed resource rather than a black box. The central goal is not simply to automate content creation, but to understand how hybrid design methods can support child-robot interactions that are more personalized, educationally meaningful, coherent over time, and still safe and more transparent.

Research Questions in 2025:

RQ1. How can a child-robot interaction be designed so that a robot can reconnect meaningfully with a child after a longer period of no contact, while preserving the sense of social continuity?

RQ2. What happens to child-robot relationship quality and memory accuracy after a long absence? To what extent does reconnecting with the robot using non-robot mediated modalities repair incompleteness, outdatedness, or errors in the robot's memory about the child, and how can this memory be repaired through interaction?

RQ3. How can children be given a more active role in inspecting, correcting, and co-constructing the robot's memory, rather than remaining passive subjects of a hidden user model?

Progress in 2025: In 2025, my work focused on reporting on my completed work, and developing and evaluating new ways of designing sustained personalized child-robot interaction. The emphasis lay on hybrid

human-AI design methods, long-term relational continuity, and child-accessible robot memory. My work also involved supervising several thesis students, working within my project lines. The main results of this work took several forms.

First, I developed the conceptual and technical foundations of a hybrid content generation framework for educational child-robot interaction, which were validated in scope of Robot Bookworm study. This work addressed how LLMs can be used not as uncontrolled generators, but as part of a constrained and human-steerable pipeline for producing child-facing interaction content. A major outcome here was a clearer articulation of hybrid design principles: AI is used selectively where it adds value, while human authors remain in control of pedagogical structure, safety, and quality oversight. This work also resulted in the AIED 2026 paper "A Hybrid Human-AI Content Generation Framework for Safe and Personalized Dialogic Learning with Children,"

Second, I developed and analyzed the Robot Bookworm line of work further. This resulted in a stronger theoretical and analytical framing around personalized dialogic book discussion, reader-book relatedness, and reading enjoyment.

Third, I developed the Reconnect / Open-Minded Robots line into a major strand of the PhD. Robot-Reconnect user study was designed and executed. This work highlighted a design gap around long-term child-robot interaction: after a long break, both the relationship and the robot's memory may have degraded, and stored child models may no longer be accurate. This shifted my focus from personalization alone toward child-in-the-memory-loop interaction. Building on this, I designed the foundations of the Open-Minded Robots study. The main result was a concrete user study design in which robot memory is treated as a co-constructed resource which children can directly inspect and influence.

Project 2.16, Anastasia Sophia Apeiron

Responsible autonomy for hybrid intelligence

Goal: We will investigate how the notion of consent can be used as an abstraction to capture and facilitate interactions among multiple humans and AI agents in realizing hybrid intelligence. We will present a formal representation of consent, what the types of consent are and how they can be inferred, and when and how an agent should request consent.

Research Questions in 2025: How do humans perceive the consent model of the agent in terms of attributed responsibility and autonomy? Does changing the consent model affect this perception or are these attributions due to other interactions?

Progress in 2025: To answer these questions, I (along with my PIs) have conducted a between-subjects vignette-style user study. We have found from 138 participants that for 1 element of responsibility and 3 elements of autonomy, a dynamic consent model is seen as resulting in a greater attribution of responsibility and autonomy, respectively. From these findings, we present three guidelines, currently under review in our IJCAI-2026 submission, that can be used to guide future design choices when building consent models for human-AI collaborative systems.

Project 2.17, Maurus Enbergs

Designing for emergence

Goal: We will develop new methodologies to design for the emergent nature of hybrid intelligent systems. We will develop formal abstractions for the analysis and design of the dynamics of hybrid intelligent systems. These will inform practical design methodologies that can enforce important emergent properties such as safety or justice in the emergent system dynamic.

Research Questions in 2025: How can control structure engineering be used to construct formal abstractions that help identify and analyze the emergence of safety hazards in real-world use cases of AI and hybrid intelligent systems?

Progress in 2025: In 2025, we conducted a series of empirical investigations into algorithmic and AI systems deployed in real-world public-sector contexts. We selected cases that exhibited complex sociotechnical interactions between AI systems, human operators, institutional procedures, and affected citizens. Using control structure engineering as an analytical lens, we examined the controls, feedback loops, and information flows present in these systems. Based on this we started to develop formal abstractions that capture not only the technical components of the AI systems, but also their interactions with organizational and human actors. We aimed to identify patterns in which algorithmic outputs can constrain human judgment in ways that reduced the broader system's ability to detect and correct harmful dynamics over time. We aim to test how the resulting abstractions can support the development of more principled design methodologies for hybrid intelligent systems. As HI systems are characterized by more complex human–machine interactions, transferring these insights is a necessary step toward addressing safety and other emergent properties, such as justice, in future HI system designs.

Project 2.19, Ludi van Leeuwen

Designing a rational process for hybrid probabilistic decision making

Goal: Our goal is to design a process for creating Bayesian Networks that can handle various real-world problems, such as the problem of the reference class, problem of the priors, causality, and others- or to show limitations of such an approach. The process should be hybrid, in that it plays to the strengths of both the BN and human modellers and users (arguers) by combining probabilistic updating rules with rational deliberation to decide which events, causalities and probabilities to apply.

Research Questions in 2025:

- What are relevant features for eyewitness testimony? How do we implement this in an agent-based model, and how do Bayesian networks model this? How do we evaluate how Bayesian networks evaluate eyewitness testimony?

- How can we create a better method for modelling complex cases in Bayesian networks?

Progress in 2025: I went to several conferences and workshops (incl. The Lorentz workshop on hybrid argumentation and responsible AI), was invited to present my work at the Girona evidence week. I worked towards an accepted ICAIL paper on reliability of testimony. I organized a workshop on evidential reasoning with AI at the JURIX in Turin.

Project 2.21, Cor Steging

Correct Reasoning in Hybrid Intelligence

Goal: The PD will investigate correct reasoning in state-of-the-art AI systems. This includes:

1) Creating benchmarks for evaluating and improving correct reasoning in large language models (LLMs).

2) Creating neuro-symbolic architectures that align data-driven and knowledge-based approaches.

Research Questions in 2025: How can we create benchmarks to evaluate reasoning in LLMs that are scalable in complexity and not prone to data contamination?

Can multi-agent deliberation be used in legal reasoning?

Can we design an Agentic Belief-Desire-Intention (BDI) model for legal reasoning?

How should we tackle sycophancy in LLMs for deliberation?

Progress in 2025: We developed parameterized, argumentation-based reasoning tasks for benchmarking the reasoning capabilities of LLMs. Since these benchmarks are dynamically generated, we avoid the issue of data contamination, and due to their scaling complexity we can continually evaluate advancing reasoning capabilities. We also evaluated multi-agent approaches to legal reasoning, and compared these against baseline LLMs, as well as developed two novel multi-agent frameworks. In addition, we have developed a framework for an agentic BDI model for legal reasoning. Lastly, with a large group of scholars, we have proposed the concept of 'reasonable parrots', where LLMs should argue with us by design in order to reduce sycophancy.

Project 2.22, Linyun Xiang

Modeling human behaviors in negotiations and deliberations

Goal: Negotiations and deliberations involve understanding and managing interpersonal dynamics, relationships, motivations, and navigating conflicts. While there has been much existing research in building negotiating agents, there remains a gap to integrate situated understanding of human behaviors and how these could affect negotiation and deliberation outcomes. What could be perceived from people's interaction and communication patterns as well as decision-making across the negotiation and deliberation process? One solution is to leverage information gathered pre- and post- interaction, as well as multimodal data during the interaction process. This project aims to develop novel frameworks and learning approaches that integrate human behavior modeling to support decision-making in real-life scenarios.

Research Questions in 2025: How can dialogue summarization be improved to capture both semantic content and emotional dynamics of multi-speaker conversations, and how can the preservation of emotional flow be effectively evaluated?

Progress in 2025: Introduced a dialogue summarization framework that models both semantic content and emotional dynamics using a hierarchical Chain-of-Agents approach. It generates topic- and participant-level summaries with inferred emotions and combines them into a final summary capturing emotional flow. Experiments show improved performance, supported by new emotion trajectory evaluation metrics.

Project 3.01, Ilke Asal

Personalization of LLMs for Promoting Effective and Long-Lasting Behaviour Change

Goal: This project aims to ethically and effectively personalize LLMs by capturing users' mental models. The goal is to deliver impactful behaviour change interventions tailored to individual needs, resulting in sustained behavioural change.

Research Questions:

1. Which behavioural determinants should be included in users' mental models for use by LLMs to personalize conversations that promote effective behaviour change?

1a. How should a mental model capturing key behavioural determinants be represented for the personalization of LLMs?

1b. What should be the most practical and efficient way to integrate this mental model into an LLM?

2. How does the integration of this mental model affect users' perceptions of the conversation and their behaviour?

Project 3.02, Tommaso Zendron

Interactive Dialogues with Logic-Based AI Systems

Goal: The general scope of this PhD project is to develop new techniques for interacting with logic-based systems (focusing on systems based on ontologies formulated in OWL), to ensure trustworthy and efficient dialogues with the user. We aim for a new form of automated reasoning that allows for iterative communication among the agents until the system provides a satisfying and tailored response to the user. We call this new reasoning task interactive abduction.

Research Questions:

RQ1: What are the theoretical foundations of dialogues with logic-based systems?

RQ2: How can we develop a prototypical generic system that allows for collaborative query answering with OWL knowledge bases?

RQ3: In what ways can user models be leveraged to minimize the friction, length, and complexity of the interaction?

RQ4: How should an ontology-driven system produce explanations that are both accurate and adaptable to the user?

Project 3.03, Johannes Bendler

Configuring neuro-symbolic HI systems with guaranteed CARE properties

Goal: The goal of this project is to develop a formal framework that links neuro-symbolic architecture design to system properties. Building on the visual Boxology language, we develop a mathematical formalization that is grounded in Category Theory, to enable proof-based guarantees of CARE properties (collaborative, adaptive, reliable, explainable). The soundness will be tested through empirical case studies of existing Hybrid Intelligence systems.

Research Questions:

RQ1: How can the Boxology framework and the CARE properties of HI systems be formalized into a shared mathematical-logical structure?

RQ2: How can properties be derived from a given neuro-symbolic architecture in a deterministic manner and how can such systems be configured to exhibit specific properties?

RQ3: To what extent does the developed framework enable the description and improvement of existing HI systems in practice?

Project 3.05, Ahmed Yesuf Nurye

Interactive robot learning with flexible human feedback

Goal: The goal of this project is to endow robots with interactive learning capabilities that can effectively incorporate diverse human input across multiple communication modalities.

Research Questions: We have formulated the following three main research questions.

1. How to effectively learn from multiple teaching signals (e.g., demonstrations, corrections, preferences) that can be expressed via various communication modalities (e.g., speech or direct manipulation)?

2. How to enable robots to reason about their own learning needs to actively query the right feedback at the right time while balancing the expected information gain with teaching effort?

3. How to enable robots to effectively account for suboptimality in human feedback arising from execution noise (e.g., imperfect teleoperation) or from limited knowledge of the optimal task solution?

I am currently working towards my first paper, which addresses the first research question.

Project 3.06, Rosie Zheng

Improving human-robot collaboration using brain-computer interfaces

Goal: We want to develop a real-time neuroadaptive system by detecting error-related potentials (ErrP) in electroencephalography data and leveraging reinforcement learning (RF) to update the robot's behaviour. By integrating neural feedback with adaptive robot learning, our research aims to enable real-time adaptation, reduce user workload, and advance intuitive, human-centered human-robot collaboration (HRC).

Research Questions: The main research question is: Can neural feedback combined with RF improve robot learning in HRC tasks? Sub-questions address (1) the state of the art of brain-computer interfaces (BCIs) in HRC through a systematic literature review; (2) the accuracy of ErrP detection during collaborative tasks; (3) whether BCI outputs can serve as a reward mechanism for real-time RL-based adaptation; and (4) the impact of neuro-adaptive collaboration on user experience, including trust and autonomy.

Project 3.08, Dmitrii (Mitja) Orlov

Knowledge Elicitation for Scientific Data in Hybrid Heterogeneous Teams

Goal: This project began in September 2025. It aims to develop and evaluate human-AI collaboration strategies for tacit knowledge elicitation. It addresses the use of Scientific Dataset reuse, where ambiguities in metadata and other descriptions are frequent and hamper reusability. We design an agentic system and interaction patterns that would benefit both the process and the outcomes of creating dataset descriptions. Our research framework seeks to establish new evaluation methods, collaborative design with domain-expert involvement, and an analysis of socio-technological tradeoffs of introducing collaborative AI systems into the practice of knowledge elicitation.

Research Questions:

1. How to evaluate the tacit knowledge representation in scientific datasets?

2. How to effectively coordinate human-AI metadata co-creation workflows?

3. How to co-design Agentic AI systems with scientific communities?

Project 3.09, Satchit Chatterji

NeSy for Safety and Norms in Multi-Agent RL

Goal: The project's core goal is to develop neurosymbolic (NeSy) approaches for multiagent reinforcement learning (MARL), where multiple learning agents interact in shared environments and must learn to cooperate while avoiding unsafe or undesirable outcomes. By combining neural learning methods with symbolic rules, norms, or formal constraints, the project aims to make MARL systems safer, more interpretable, and better aligned with human governance. A central focus is understanding how these symbolic constraints shape emergent agent behavior, how well they match human interpretations of policies or norms, and how such systems can support Hybrid Intelligence use cases, especially in legal or governance-related settings.

Research Questions: The primary research question is as follows:

How can we develop neurosymbolic systems that ensure safe and effective multiagent cooperation under human governance (e.g., policymaking)?

This research question can further be decomposed into two research questions:

RQ1) To what extent can we develop a neurosymbolic system that can guarantee safety and agent-agent cooperation?

RQ2) How does the employed formal language impact the alignment between emergent collective behaviour of the agents and the understanding of the human agent (e.g., a policy maker) on the semantics of formal constraint?

Project 3.10, Selina Khan

Interactive Exploration of Bias in Cultural Heritage Collections

Goal: The goal of this project is to develop interactive AI methods that enable the exploration and analysis of bias in large-scale visual cultural heritage collections without relying on costly or sensitive manual annotations. It aims to leverage human-AI interaction to combine domain expertise with adaptable vision-language models, allowing users to uncover nuanced patterns of representation and bias. A central objective is to design systems that can learn from real-time user feedback and personalize their behavior to support diverse investigative perspectives.

Research Questions: How can we use interaction to overcome the reliance on labels for the analysis of bias in visual cultural heritage collections?

Project 3.11, Melika Davood Zadeh

Interpretable and checkable communication of natural language decisions

Goal: We aim to develop a system that can process scientific documents efficiently, make decisions based on the context and generate verifiable rationale or insights for the decision of the model. The task performed by the model can vary from reviewer paper matching, document correlation and similarity, QA and other related document-based tasks. Further, the extracted rationals can be presented to the user and be evaluated how good and informative the provided insights are, and to be used to provide feedback to the system.

Research Questions:

RQ1: How should rationales be selected in order to best assist the human expert's performance at their task?

RQ2: What algorithms or assumptions are needed to effectively and efficiently extract evidence from long documents? How much context is necessary for verifying a decision?

Project 3.13, Charlotte Ning

Multimodal modeling of trust for trust calibration

Goal: The goal of this project is to equip an interactive agent with the capability to (1) infer the user's trust state towards the agent during the interaction, and (2) proactively calibrate this trust state when it deviates from an appropriate level. To enable online and unobtrusive trust measurements, this project explores the integration of multimodal large language models (MLLMs) and think-aloud protocols.

Research Questions: How can AI agents infer and proactively calibrate human trust during interaction using think-aloud-informed multimodal modeling?

Project 3.16, Ruben Weijers

Human-Robot Interaction strategies for group collaboration

Goal:

- Find out if solution-focused interaction strategies negatively influence basic psychological needs in humans during group interaction with a robot
- Train a capable robot when and how to provide solutions from data in group collaboration
- Leverage scaling AI capabilities to improve group collaboration in humans through understanding of needs and goals

Research Questions:

- What is the effect of robot-provided solutions on the interaction topology of group conversation?
- What is the effect of robot-provided solutions on participants' autonomy, competence, and relatedness?
- What is the effect of context-based robot-provided solutions on creative task outcomes in group collaboration?
- How do we train a system to adapt to a group to better foster group collaboration leveraging both the AI's and human's unique capabilities?

Project 3.17, Anuj Singh

Causal Representations for Open-Ended, Lifelong Hybrid Intelligence

Goal: The goal of my PhD is to develop lifelong learning, socially perceptive agents that learn causal representations of both their environments and their interaction partners in multi-agent settings. Rather than relying solely on static, internet-scale datasets used for pre-training, these agents should be capable of continual in-situ adaptation at test time during deployment.

This requires integrating causal representation learning, partner modelling, and continual/open-ended learning into a unified framework for adaptive social intelligence, enabling agents to reason about changing environments, human behaviour, and collaborative tasks as they unfold.

Research Questions:

RQ1: How can agents learn causal representations of the environment and of other agents' mental states?

I will study how agents can learn latent causal models of world dynamics and partner beliefs, desires, and intentions from interactive experience. These representations are intended to support counterfactual reasoning, adaptive planning, few-shot generalization, and the retrieval of task-relevant memories in evolving environments.

RQ2: How can causal social agents continually improve in open-ended, lifelong learning environments?

I will investigate how agents can continually adapt during deployment by distinguishing stable causal mechanisms from those that shift across tasks, environments, and interaction partners. This includes studying directed exploration through intrinsic motivation, mechanism-level adaptation, and knowledge sharing across specialist agents to support robust learning in open-ended social settings.

Project 3.22, Berke Yazan

Justifications in human-AI teams

Goal: In hybrid teams, behavior is coordinated through norms that are designed to promote particular values. As teams grow larger or operate over longer time spans, norms can become outdated, can conflict with one another, or can be poorly suited to the situation at hand, and strict compliance can fail the values that the norms were meant to serve. In such cases, deviation can be the more responsible and accurate option. This project studies when an agent in a human-AI team is permitted to deviate from a norm, and how it can justify that deviation in terms of the norm's own intent. The main focus is also on the justification itself, including what makes a justification sufficient, how it should be communicated to other team members, and how it supports trust between humans and agents when the prescribed action is not the one taken. The aim is a normative framework and a deliberation architecture that together describe when deviation is justified, how the justifications are produced, and how the team learns from them over time.

Research Questions:

1. Under what conditions is it justifiable for an agent in a human-AI team to deviate from a norm, particularly in larger teams where multiple norms may apply at once and conflict with one another?

2. What makes a justification for such a deviation sufficient, and how does it affect trust, accountability, and learning among human and artificial team members?

3. How do different stances towards norm authority, from strict compliance to value-led deviation, affect the behavior and outcomes of the team as a whole?

Project 3.23,Debargha Ghosh

"Causal insights": adaptive explanations for interpretable probabilistic models

Goal: This project aims to incorporate causality into explanations of probabilistic models with the purpose of providing meaningful explanations in a hybrid intelligence context where humans interact with an AI system that helps them understand the effects of their actions and how to improve upon them.Here, we will introduce the concept of causality into the explanation of PGMs, by exploiting knowledge about causality (or lack thereof) taken into account in the construction of the networks, as well as in the computations performed for determining the explanations. Since the existence of true causal relations can often not be determined with certainty, we will design methods for establishing and explaining how robust the network predictions are to varying causal assumptions. We aim at explanations that have desirable properties, e.g., being contrastive, selective and interactive and adapted to the needs of different types of users. In addition, we aim to develop metrics for assessing the quality and robustness of explanations, especially in relation to the users' needs.

Research Questions:

1. Effect of assumption violations on causal discovery and downstream inference.

2. Robustness of causal discovery models

3. Causal aware adaptive explanations for Bayesian Networks**.**

Project 3.25, Tom Kouwenhoven

Social Cognition for Hybrid Collaborations

Goal: This project revolves around “machine psychology” in the context of social cognition. Its aim is to benefit human-machine collaboration by unravelling whether and how machines differ from humans in situations that require social cognition. Specifically, we aim to understand which social mechanisms are needed for successful human-machine interaction, when they matter, and how they operate. The project centres on Theory of Mind (ToM) as a proxy for social cognition, but also considers a broader set of mechanisms and strategies, ranging from low-level, implicit processes grounded in situated and embodied cognition, to higher-level reliance on norms, conventions, and common ground.

We take two complementary approaches. First, we study and train contemporary Large Language Models while carefully manipulating their input data, looking for “traces of social cognition” in the resulting behaviour. Second, we deploy computational agents (LLM-driven and otherwise) in experimental settings where humans and machines collaborate. Rather than assuming particular mechanisms underlie social behaviour, we adopt a pragmatic stance toward “understanding” and focus on what can be achieved through collaboration and alignment in practice. This perspective contributes to our understanding of artificial agents as active participants in society.

Research Questions:

1. Which perspective-coordination mechanisms are available to machines in collaborative settings, and how do these compare to the mechanisms humans rely on?

2. What factors (e.g., model architecture, training data, or task design) determine whether machines succeed on tasks targeting specific subskills of social cognition (e.g., false-belief reasoning, common-ground tracking, norm inference)?

Project 3.28, Mireia Yurrita Semperena

Methodology and Metrics for Hybrid Intelligence

Goal: The goal of this research project is to scientifically investigate Hybrid Intelligence and to contribute toward establishing it as a research field. The focus of the project is on the development of a methodology to identify different dimensions that characterize hybrid intelligence, design metrics or processes to measure or assess such dimensions, and create use-cases to demonstrate different aspects of hybrid intelligence. The developed methodologies are expected to be used to evaluate how well human-AI systems provide solutions for hybrid intelligence, and to facilitate a hybrid design of new human-AI systems.

Progress in 2025: Developing Guidelines for Human-LLM Agent Teams: A Multi-Stakeholder Lens

Abstract: Agents based on Large Language Models (LLM agents) have the potential to work with humans as part of a team to achieve specific goals. The natural language interface of LLM agents and their high level of autonomy enables more seamless collaborations than previous technologies, allowing them to carry out tasks autonomously and engage in conversations with humans, e.g., to clarify goals, request authorizations, or double-check decisions. However, the current literature lacks systematic design guidelines for these human-LLM agent teams. This gap might foster misunderstandings, misuse of autonomy, and lack of common ground, potentially leading to collaboration pitfalls. To mitigate these risks, we develop 24 guidelines for the principled design of human-LLM agent teams. We adopt a multi-stakeholder approach and propose guidelines for LLM agents, human team members, team designers and embedding organizations. To develop these guidelines, we distill design recommendations from an exploratory workshop with 15 experts on human-AI teaming and a literature review of 93 empirical papers in human-LLM collaboration. Drawing from literature on human teams, we conceptually categorize the recommendations across different stages of the teaming process. A user study with 10 additional experts suggests the guidelines can help prevent collaboration pitfalls in human-LLM agent teams within workplace settings.

Yurrita, M., Dell’Anna, D., Murukannaiah, P. K., Jonker, C. M., & Yolum, P. (2026). Developing Guidelines for Human-LLM Agent Teams: A Multi-Stakeholder Lens. In AAMAS 2026- Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (pp. 1956-1966). Association for Computing Machinery (ACM).

Project 3.30, Panos Aronis

Learning Trade-offs in Multi-Stakeholders Multi-Objective Problems

Goal: Societally relevant problems come with a myriad of challenges involving complex decision-making processes that must balance multiple conflicting objectives and coordinate the actions of various decision-makers. In other words, many real-world decision-making problems are multi-objective multi-agent in nature. Think for example about social infrastructure management and optimisation in urban settings. Decisions in this case should yield efficient use of available resources (e.g., land, capital), ensure equity and accessibility to amenities, while adhering to sustainability and resilience principles. A trade-off solution is required

among these interdependent and potentially conflicting objectives, that reconciles the perspectives and preferences of the multiple involved stakeholders (e.g., local, regional or national governmental institutions and regulatory bodies, private contractors, residents of different age groups, community organisations). This challenge is ubiquitous, with other examples including managing traffic systems, electricity grids, supply chains, epidemiology, and so on.

This project aims to develop the first user-centred pipeline for decision-support systems in multi-agent multi-objective settings. This pipeline will bridge the gap between learning and solution selection, helping stakeholders understand trade-offs and align decisions with their preferences. This is an interdisciplinary project that combines multi-objective optimization, multi-agent reinforcement learning, game theory, and computational social choice.

Research Questions:

- How can we develop decision support systems that support human collaboration?
- How do we navigate and ensure preference reconciliation in a polyvocal complex decision support scenario?

Project 3.31, Ahdi Alipour

Designing Inclusive and Responsible AI Agents for Diverse User Groups

Goal: Our project investigates how key human factors, such as age, gender, cultural background, personality, and especially digital literacy, shape interactions with AI agents. Among these factors, digital literacy plays an important role in enabling engagement with AI agents, yet it remains underexamined in past research. Our main goal is to understand how different levels of digital literacy affect trust in AI agents, particularly in terms of trust calibration. Under-trust may lead to avoidance of beneficial technologies, while over-trust may result in over-reliance on flawed outputs. Since developers often overlook the diversity of human users, these risks particularly affect marginalized and underrepresented groups, such as older adults, people with disabilities, individuals from lower socioeconomic backgrounds, and refugees. Our project therefore aims to contribute to the design of more inclusive AI agents that accommodate diverse user capabilities in public services.

Research Questions: How can we design an AI agent that adapts responsibly to different users, considering their digital skills, personality, age, culture, and gender?

Project 3.32, Joséphine Mélot-Chesnel

Designing Trustworthy Generative AI Agents for Collaborative Decision Making

Goal: This project examines how contextual factors (e.g., task difficulty, time pressure and collaboration setting) and agent characteristics (e.g., explanation strategy, display of uncertainty and perceived competence) shape both subjective trust and observable trusting behaviors (e.g., reliance, verification, delegation) in human-genAI collaboration. The overall aim is to advance human-centered AI design and deepen understanding of how trust is formed, maintained, and calibrated to mitigate both under-trust and over-reliance in human–genAI decision-making environments.

Research Questions: How is human trust formed, maintained, and calibrated when collaborating with generative AI agents in decision-making tasks?

Project 3.33, Sayak Mukherjee

Data-Efficient Machine Learning for Context-Sensitive Affective Computing

Goal: Our research aims to develop context-aware affective computing systems capable of dynamically inferring human emotional states from multimodal signals. A central challenge is that emotions are strongly context-dependent, i.e., the same expression or behaviour can convey different emotions depending on situational, cultural, and personal factors. Rather than relying on labelled datasets, which are constrained by ethical and privacy concerns and often yield models that learn superficial correlations, we frame emotion recognition as an ill-posed inverse problem. Inspired by how humans use structured knowledge about events, goals, beliefs, and situations to interpret emotion, our work aims to leverage causal models grounded in the intuitive theory of emotion and combine them with large multimodal foundation models to build affective systems that are generalisable across contexts, data-efficient in adaptation, and interpretable in operation.

Research Questions:

1. How robust are existing end-to-end emotion recognition models to contextual variation?
2. How to reason about emotion from context using LLMs that have an intuitive theory of emotion?
3. How can causal models of emotion be combined with multimodal LLMs to integrate emotional cues from context and expression?

4. How can context-aware emotion recognition systems support planning and decision-making in real-world applications?

Project 3.35, Sien Jansen

Formalizing ethical use of software components to mitigate bias in automated decision making

Goal: The aim of this project is to develop a formal language and a practical methodology that supports reasoning, verification, and certification of fairness in software systems that use AI. This language will be based on both deontic logic and description logic.

Research Questions:

RQ1 How can fairness within software be formalised using logic-based formalisations?

RQ2 How can a logical formalism combining description logic and deontic logic be used to verify fairness within a software system and its components?

RQ3 In what way can such a framework assist users in selecting an appropriate fairness definition for a software system across varying domains and systems?

RQ4 How can such a framework generate formal certificates to support bias mitigation?

Project 3.38, Sebastiaan De Peuter

Optimal Explanations in Hybrid Teams

Goal: This project investigates how to capture and maximize the value of explanations from an AI within a hybrid human-AI team. Explainable AI has long been touted as a way to help humans form a better understanding of an AI, allowing them to better decide when and how to rely on AI support. However, in practice there is little evidence that explainable AI has a significant effect on human-AI team decision quality. A central challenge is that there is a lack of theory around what makes explanations helpful. This project therefore looks to formally model explanations within the context of a human-AI hybrid team. The basic insight is that by treating explanations as communication influencing the beliefs of the human, we can quantify their information content and how much they improve future team decisions. Furthermore, we can then use this formalization to allow an AI to learn which explanations are most helpful to a specific human team mate.

Research Questions:

RQ1: How can the effect of an explanation be formalized within the human-AI team? How do we formalize explanations as communication influencing a human's beliefs, so that we can capture their information-theoretic and utility value?

RQ2: Building on the formalization from Q1, how can an AI system optimally explain? Can we treat it as a decision problem so that the AI can learn to optimally explain?

RQ3: How can AI adapt its explanations to a human team mate? Can we optimize explanations by adapting them to the current knowledge and beliefs of the human?

Hybrid Intelligence Management Team

Management Team

Prof. Catholijn Jonker

TU Delft

Prof. Frank van Harmelen

Vrije Universiteit Amsterdam

Dr. Erman Acar

Institute for Logic, Language and Computation (ILLC)
University of Amsterdam (UvA)

Dr. Wendy Aartsen

Hands4Grants

Dr. Nazmiye Ozkan

Hands4Grants

The Hybrid Intelligence consortium gratefully acknowledges the support of the Gravitation Programme of the Netherlands Ministry of Education, Culture and Science (OCW) and the Netherlands Organisation for Scientific Research (NWO).