DAGSTUHL
REPORTS

Volume 13, Issue 11, November 2023

ISSN 2192-5283

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at
https://www.dagstuhl.de/dagpub/2192-5283

Publication date
April, 2024

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at `https://dnb.d-nb.de`.

Digital Object Identifier: 10.4230/DagRep.13.11.i

Aims and Scope
The periodical *Dagstuhl Reports* documents the program and the results of Dagstuhl Seminars and Dagstuhl Perspectives Workshops.
In principal, for each Dagstuhl Seminar or Dagstuhl Perspectives Workshop a report is published that contains the following:
- an executive summary of the seminar program and the fundamental results,
- an overview of the talks given during the seminar (summarized as talk abstracts), and
- summaries from working groups (if applicable).

This basic framework can be extended by suitable contributions that are related to the program of the seminar, e. g. summaries from panel discussions or open problem sessions.

Contact
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
Dagstuhl Reports, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
reports@dagstuhl.de
https://www.dagstuhl.de/dagrep

Report from Dagstuhl Seminar 23451

Visualization of Biomedical Data – Shaping the Future and Building Bridges

Katja Bühler[*1], Barbora Kozlíková[*2], Michael Krone[*3], Cagatay Turkay[*4], and Ramasamy Pathmanaban[†5]

1 **VRVis – Wien, AT.** katja@vrvis.at
2 **Masaryk University – Brno, CZ.** kozlikova@fi.muni.cz
3 **Universität Tübingen, DE.** michael.krone@uni-tuebingen.de
4 **University of Warwick – Coventry, GB.** cagatay.turkay@warwick.ac.uk
5 **Ghent University, BE.** Pathmanaban.Ramasamy@UGent.be

Abstract

The last decades of advancements in biology and medicine and their interplay with the visualization domain proved that these fields are naturally tightly connected. Visualization plays an irreplaceable role in making, understanding, and communicating biological and medical discoveries. The goal of Dagstuhl Seminar 23451 was to serve as an interdisciplinary platform for a collective approach to the contemporary and emerging future scientific and societal challenges at the intersection of visualization, biology, and medicine in the context of increasing complexity in data, data analytics, and data-intensive science communication. Building on the success of the previous seminars and our ongoing community efforts, participants of this seminar critically tackled highly relevant scientific questions of interest to the bioinformatics, medical informatics, and visualization communities. These challenges include the increasing complexity and amount of data that are produced in biomedical research, the role of visualization in supporting interdisciplinary research and in communicating biological and medical discoveries to experts and broader audiences, and visualization for a user-centric and trustworthy explainable AI in biomedical applications. The seminar was an important step towards strengthening and widening a sustainable and vibrant interdisciplinary community of biological, medical, and visualization researchers from both academia and industry through an in-depth, comprehensive, and inclusive exchange of ideas, experiences, and perspectives. The identified key topics span methodological, technical, infrastructural, and societal challenges. The discussions and exchange of ideas revolved around the most pressing problems among the biological and biomedical domains and how these problems could be approached through data visualization, thus opening up room for innovation in designs and methodologies.

Seminar November 5–10, 2023 – https://www.dagstuhl.de/23451

2012 ACM Subject Classification Applied computing → Bioinformatics; Applied computing → Health informatics; Human-centered computing → Scientific visualization; Human-centered computing → Visual analytics; Human-centered computing → Information visualization; Human-centered computing → Visualization theory, concepts and paradigms; Human-centered computing → Visualization design and evaluation methods; Computing methodologies → Neural networks

Keywords and phrases biology, computational biology, interdisciplinary, medicine, visualization

Digital Object Identifier 10.4230/DagRep.13.11.1

* Editor / Organizer
† Editorial Assistant / Collector

Visualization of Biomedical Data – Shaping the Future and Building Bridges, *Dagstuhl Reports*, Vol. 13, Issue 11, pp. 1–19
Editors: Katja Bühler, Barbora Kozlíková, Michael Krone, and Cagatay Turkay
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Barbora Kozlíková (Masaryk University – Brno, CZ)
Katja Bühler (VRVis – Wien, AT)
Michael Krone (Universität Tübingen, DE)
Cagatay Turkay (University of Warwick – Coventry, GB)

The goal of this seminar was to tackle pivotal challenges concerning the future of visualization for biology and medicine. Emphasizing a collaborative and interdisciplinary approach, the seminar brought together 36 leading experts and emerging researchers from academia, media, and industry covering the fields of bioinformatics, biological, and medical visualization. The seminar endeavored to delve into the multifaceted role of biomedical visualization in science communication and interactive exploratory analysis. This involves navigating the increasing complexity of biomedical data that can be acquired today, with a spotlight on integrating information from diverse sources and modalities. Examples are visualizing multi-omics data, facilitating within- and across-cohort analysis, and illustrating dynamic cellular processes. Additionally, there is an intersection of visualization with artificial intelligence in biomedical visual analytics applications, exploring how to make Artificial Intelligence (AI) outcomes comprehensible and actionable in critical decision-making processes. The seminar focused on four key topics vital to the intersection of visualization, biology, and medicine:

- the role of biomedical visualization in communicating complex processes to both expert and broad audiences,
- the challenge of mastering the complexity and multi-modality of biomedical data, emphasizing integrative approaches that can help bridge various scales and data modalities from diverse sources,
- the synergy between visualization and AI in biomedical applications, and
- shaping curricula for biological and medical visualization and the definition of content, structure, and goals for a future educational platform.

During the first day of the seminar, these topics were further developed through joint brainstorming, taking into account the interests of the participants and the need to keep up with the fast-paced developments in the field of AI and its increasing influence on biomedical visualization and related research fields. Six working groups emerged from this process, working over the next four days of the seminar accompanied by daily summary sessions in the plenum:

1. *Bridging Scales, Data, and Modalities* to discuss the changing nature of biological visualization adapting to the needs of biologists as a result of the massive increase in the scale of the data and the increasing need to cross traditional domain boundaries.
2. *The Role of Data Visualization in Instigating Behavior Change (to Promote Healthy Lifestyles)* to discuss how data visualizations can be used in a dialogue between medical practitioners, researchers, and the public as a fundamental component in preventive healthcare.
3. *A Rollercoaster Ride into the Future: AI-in-the-loop for Visual Workflows in Biomedical Data Analytics* to discuss opportunities and challenges that recent developments in AI present for the visual analysis of biomedical data.
4. *Components of a Syllabus for Life Science Data Visualization* to discuss topics related to education and educational platforms.

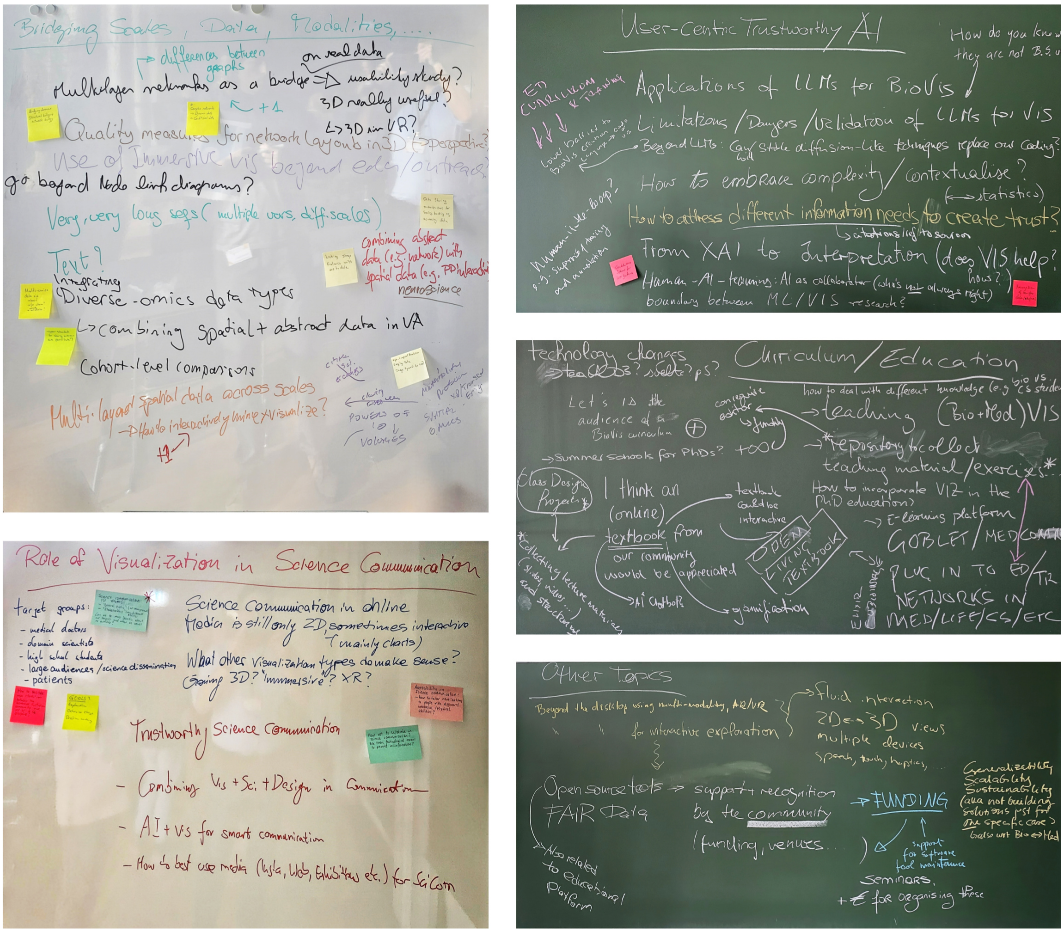


Figure 1 Results of the brainstorming session on Day 1.

5. *Beyond the Desktop: Leveraging Immersive Environments for Biomedical Data Analysis – Challenges, Vision, and Guidelines* to discuss challenges in immersive analytics in biomedical analysis scenarios.
6. *Spatio-Textual Interaction in Visualization* to experimentally explore the capabilities of current large language models for spatial interaction with data.

Five lightning talks complemented the group work and acted as inspirational prompts to spark discussions:

- *Exploring Relations among Topics in Neuroscience Literature using Augmented Reality*, by Lynda Hardman, CWI – Amsterdam, NL & Utrecht University, NL
- *Is that right? Visualizations for scientific data quality control*, by Devin Lange, University of Utah
- *Can ML/AI be Taught in Schools?* by Blaz Zupan, University of Ljubljana
- *Visualization building blocks for analysis, not the end of pipelines*, by Trevor Manz, Harvard Medical School
- *How to design data visualizations for a (very) broad audience*, by Matthias Stahl, Der Spiegel

The seminar stimulated lively discussions on the future of biomedical visualization research and education in response to the increasing data complexity and related demands on interactive data analytics systems and the impact of AI on our field. All working groups are planning follow-up activities, including meetings and joint publications based on the insights gained.

Overall the seminar was a great experience bringing together researchers from different academic and non-academic backgrounds, experience and interests bridging from life science to medical applications to communication and media and from visualization technology to display technology and recent development in AI. We envision that the outcomes from the working groups will foster the links between these areas and help establish a consolidated research agenda to approach the challenges that lie ahead.

2 Table of Contents

3 Overview of Talks

3.1 Exploring Relations among Topics in Neuroscience Literature in Augmented Reality

Lynda Hardman (CWI – Amsterdam, NL & Utrecht University, NL)

Maintaining an overview of publications in the neuroscientific field is challenging, in particular in tasks such as investigating relations between brain regions and brain diseases. To support neuroscientists in this challenge, we investigate whether using Augmented Reality can make analyses of literature more accessible and integrate them into current work practices. We explore a number of questions, such as whether interaction with a large body of literature using topics provides a useful way for neuroscientists to explore and understand specific relationships. Our assumption is that by providing overviews of the correlations among concepts, these will allow neuroscientists to better understand the gaps in the literature and more quickly identify suitable experiments to carry out. We currently provide functionality to visualize and filter direct and indirect relations and to compare the results of queries. Our visualization work is based on an analysis of the neuroscience publications in PubMed. This provides an association graph among topics involving cognitive functions, genes, proteins, brain diseases and brain regions. We describe our prototype 3D AR implementation DatAR and challenges we face.

3.2 Is that right? Data visualization for quality control

Devin Lange (University of Utah – Salt Lake City, US)

Data quality control does not always excite visualization researchers. But why? High quality data is critical for high quality research. In this case, quantitative phase imaging is being explored for improving the creation of treatment plans for cancer patients. We developed Loon, a visualization system that uses exemplars to combine different data sources to aid in data quality control.

3.3 Visualization building blocks for analysis, not the end of pipelines

Trevor Manz (Harvard University – Boston, US)

Many visualization tools do not directly integrate into popular computational notebook environments. This introduces an overhead to visualization, delegating purpose-buit visualizations for the end of pipelines rather than integrating within familiar workflows. We developed anywidget, an open-source toolkit to simplify extending Jupyter notebooks with custom interactive visualizations. We demonstrate the use of our toolkit to integrate a genome browser, HiGlass, into computational notebooks to enable new interactive analysis workflows.

3.4 How to design data visualizations for a (very) broad audience

Matthias Stahl (DER SPIEGEL – Hamburg, DE)

Broad audiences are hard to define and have multidimensional interests. However, a well characterized audience is necessary to design and craft effective data visualizations. During my journey in data journalism and visualization in the newsroom of DER SPIEGEL, I got to know two simple tricks how to approach this dilemma.

1. **Show the raw data**
2. **Tell at least one story**

1. **Show the raw data**
   In many cases it's a good choice to not show aggregated data like means, medians and standard deviations. It is more feasible to show the original data than the more abstract and unemotional aggregates see Fig. 2.

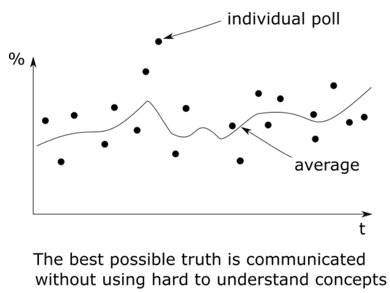


Figure 2 Poll of polls showing the polls of every institute.

2. **Tell at least one story and explain a lot as shown in Fig. 3**

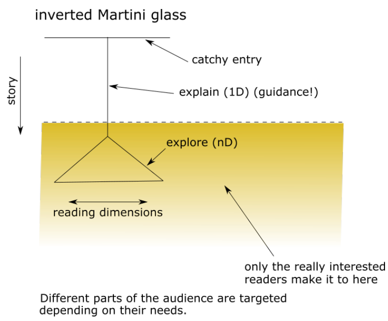


Figure 3 Inverted martini glass approach to explain a story.

3.5 Can ML/AI be taught in schools?

Blaz Zupan (University of Ljubljana, SI)

We should democratize teaching and knowledge about machine learning to the point where children in schools understand conceptually what is happening. I present a case of analyzing the positional (geo) data of surnames in Slovenia and how we trained over a thousand kids from primary and secondary schools in machine learning and data literacy using this data. The training included workshops for teachers, the production of video introductions for kids, and the design of quizzes of other written material (see `https://orangedatamining.comandhttp://pumice.si/en`).

4 Working groups

4.1 A Rollercoaster Ride into the Future: AI-in-the-loop for Visual Workflows in Biomedical Data Analytics

Thomas Höllt (TU Delft, NL), Jan Aerts (Amador Bioscience – Hasselt, BE), Marc Baaden (Laboratoire de Biochimie Théorique – Paris, FR), Stefan Bruckner (Universität Rostock, DE), Katja Bühler (VRVis – Wien, AT), Mennatallah El-Assady (ETH Zürich, CH), Zeynep Gümüs (Icahn School of Medicine at Mount Sinai – New York, US), Tobias Isenberg (INRIA Saclay – Orsay, FR), Renata Georgia Raidou (TU Wien, AT), Timo Ropinski (Universität Ulm, DE), Thomas Schultz (Universität Bonn, DE), and Pere-Pau Vazquez (UPC Barcelona Tech, ES)

We discussed the opportunities that recent developments in artificial intelligence (AI), including foundation models, generative AI, and an increasing ability to integrate multimodal data, present for the visual analysis of biomedical data, and what novel needs for visual interaction might arise from an increased use of such techniques within biomedical workflows. In particular, we covered the following aspects:

- Large language models (LLMs) and other multimodal foundation models facilitate the translation of domain-specific questions and requirements into the generation, modification, and interpretation of visual representations, and might contribute to more user-adaptive and contextualized visualization systems. Initial proof-of-concepts exist to comprehensively support the visual analytics workflow through LLMs. However, they are currently limited to individual and relatively small datasets, while biomedical applications typically require the analysis of large, complex, heterogeneous, and interconnected datasets.
- In a specialized subgroup during our breakout session, we delved into the potential applications of LLMs in comprehending and replacing various aspects of data visualizations, including spatial inputs, operations, semantics, and contextual information. The versatility of LLMs has made them suitable for a wide range of natural language processing tasks, with a notable example being ChatGPT. To this end, we conducted a preliminary

investigation of the usefulness of conversational interfaces provided by ChatGPT to steer users through data visualizations, extract insights, and potentially enhance or substitute visualization components. We specifically investigated the capacity of ChatGPT to alter the conceptual (e.g., data aspects), spatial (e.g., layouting or positional relationships within the visualization canvas), or semantic (e.g., meaning and interconnections between data elements) context of visualizations. A pilot study employing a simple scatterplot (2D+color) illuminated how ChatGPT could unveil or emphasize patterns and selections within the data, such as instructing to "highlight the green data points with a value of feature x higher than the average" or "order the blue data points from left to right and bottom to top" or "highlight the data points that have been selected within the rectangle." Our preliminary findings suggest that integrating LLMs into Visual Analytics solutions holds promise for generative AI-driven Visual Analytics and guiding users through data visualizations. Looking ahead, it would be intriguing to explore whether conversational interfaces can contribute to supporting, enhancing, or altering the interpretability of visualizations. Subsequent research could delve into more nuanced interactions with complex datasets, such as investigating visualization semantics (e.g., ontologies, visual metaphors, and abstractions), exploring user feedback mechanisms for iterative improvements (e.g., semantic zooming) and the provenance of such mechanisms, or scrutinizing further language-related semantics (e.g., labeling, descriptions, and textual elements). Finally, it is imperative to consider the ethical implications associated with relying on LLMs to shape data visualization experiences.

- Machine learning approaches are already established in practice for specific subtasks, such as dimensionality reduction of -omics data, or medical image segmentation. However, the limitations of these methods, especially when trained with the limited amount of data that is available in many biomedical applications, are still not sufficiently understood. For example, most deep learning techniques that are used today do not provide an accurate indication of the uncertainty in their estimates, and cannot detect whether a given input is sufficiently similar to their training data to be reliably processed. Even after the deployment of automated systems, continuous quality control remains essential, and is best facilitated by suitable visual interfaces.
- Explaining the predictions and outputs of machine learning methods is not only helpful for the development of such methods, but is often also a requirement for integrating them into scientific workflows (which ultimately aim for mechanistic understanding, not just predictive power) or clinical decision processes that require human experts to weight evidence from various data sources. However, the computations within machine learning methods and available model level explanations do not usually map directly to human mental models. Therefore, it is non-trivial to design explanations in such a way that humans will actually understand why the model generated a specific output, as opposed to rationalizing the proposed explanation in terms of their own mental model. We believe that future explanations should go beyond the attribution maps that are commonly presented in current work, and should combine multiple modalities, such as images and text.
- How to communicate model outputs in a way that experts and, where applicable, also the general public, will make rational and informed decisions based on them, has still not been sufficiently studied.

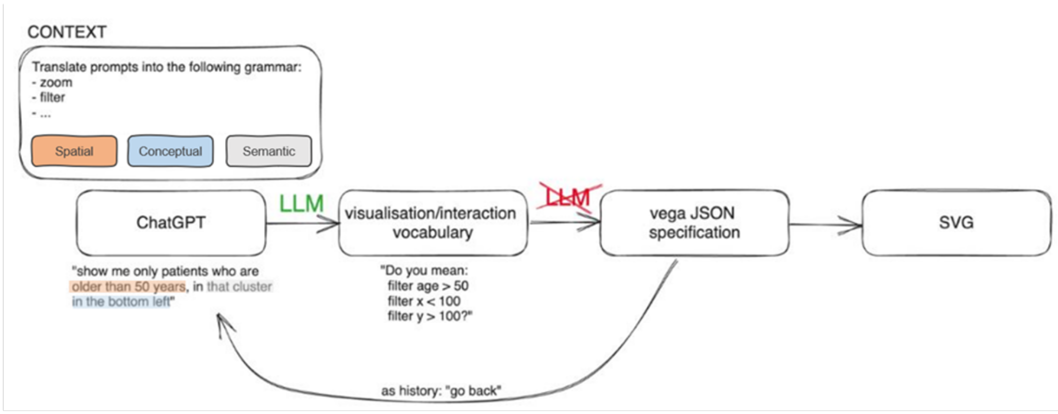


Figure 4 A model for involving large language models in visualisation development process.

4.2 Components of a syllabus for life science data visualization

Barbora Kozlíková (Masaryk University – Brno, CZ), Jan Aerts (Amador Bioscience – Hasselt, BE), Marc Baaden (Laboratoire de Biochimie Théorique – Paris, FR), Helena Jambor (Universitätsklinikum TU Dresden, DE), Georgeta Elisabeta Marai (University of Illinois – Chicago, US), Kay Katja Nieselt (Universität Tübingen, DE), James Procter (University of Dundee, GB), Renata Georgia Raidou (TU Wien, AT), Matthias Stahl (DER SPIEGEL – Hamburg, DE), and Blaz Zupan (University of Ljubljana, SI)

Abstract

The working group on building the basic components of a syllabus for teaching life science data visualization to diverse target audiences was formed at the beginning of the seminar and collaborated throughout the whole week. The group participants came from diverse fields (both in academia and industry), which made the discussions very productive. The main objectives and goals of the working group that were set up at the very beginning can be summarized as:

- Recapitulation of past syllabus planning activities and attempts at building the syllabus;
- Discussion about the immediate steps that can lead to a tangible output of these efforts;
- Preparation of a prototype syllabus, including learning outcomes, and drafting an educational publication that summarizes the challenges and recommendations we identified;
- Setting up a repository for the participants of the seminar to share teaching materials in the context of the syllabus.

Recapitulation of past activities and discussion of the next steps

Within the BioVis Dagstuhl Seminar series, where this seminar is already the fourth one, participants addressed the issue of missing guidelines and syllabi for teaching visualization in life sciences several times. There were two main outputs coming from these efforts: the categorization of the existing techniques and methods into a complex table (see Dagstuhl Seminar no. 21401 report), structured according to the scale (spanning from atoms through tissues to populations, see Fig.5).

Figure 5 Bio-medical scales. Image courtesy of Jan Byska and Noeska Smit.

The table with techniques and methods then formed the basis for the second initiative, the Spring School on Biomedical Data Visualization (`https://biomedvis.github.io`), where experts from selected fields prepared a talk on a given topic, derived from the table. Although this school has been running already for three years (`https://biomedvis.github.io/2021/`, `https://biomedvis.github.io/2022/`, `https://biomedvis.github.io/2023/`) and collects study materials that were then made freely available, it is still lacking the systematic construction of a syllabus for teaching life sciences visualization that can be shared and followed.

Thus, within this Dagstuhl Seminar, we decided to focus specifically on the challenge of creating such a syllabus. We did a first pass through the literature search to get inspiration from other disciplines or from teaching visualization in general [1]. We identified the main target audience of the teaching activities, and then used that information to determine the direction of the syllabus construction. We summarized the expected core competencies and minimal prerequisites, and then we prepared the target learning outcomes and the syllabus itself.

Learning outcomes

The **learning outcomes** and therefore core competencies are modeled using the terminology and concepts of Bloom's taxonomy [2]. When designing the syllabus, the following overall learning objective was considered: ***General DataVis literacy*** with a focus on bio-medical data. The students should gain the skills to read, analyze, and understand the visual representations used in biomedical domains. By these skills, they **understand** the relationship between visual analysis and the application domains from biology and medicine. Furthermore, students will be able

- to use the principles of human perception and cognition in visual biological and medical data analysis;
- to understand and use visual design principles;
- to know the basics and do's and don'ts of visualization (including best practices);
- to critically evaluate visual representations of bio-medical data and suggest improvements and refinements;
- to apply a structured design process to create effective visualizations;
- to create low-level prototypes for bio-medical data visualizations;
- to create simple interactive (web-based) visualizations;
- to communicate visualizations (orally or written).

The **syllabus** consists of the following elements:

1. Introduction: What is data visualization; Why do it; History
2. Data types in vis (as abstractions, using bio examples)

3. Color and perception
4. Marks and channels
5. Visual design principles and layouts
6. Visual scalability (as abstraction, using bio examples)
7. (Biology) data types
   a. Genes and Genomes
   b. Omics (quantitative data)
   c. Phylogenetic Trees and Hierarchies
   d. (Biological) Networks
   e. Molecular Structures (3D and 2D abstraction)
   f. Images (Medical images, Light and Electron Microscopy images, Gels and Plates, photos)
8. Interaction and faceting (e.g., brushing and linking)
9. Low-fidelity Prototyping
10. High-fidelity Prototyping (e.g. with Observable)
11. Evaluation
12. Ethics
13. Data-Driven Storytelling
14. BioVis software tools (Circos, Cytoscape etc.) and critique

Preparation of a publication

Already within the seminar week, the participants started to sketch the first version of the educational publication and discussed potential publication venues (for example, PLOS Computational Biology Education). Therefore, finalizing and submitting this publication will be one of the main priorities after the seminar.

Setting up a teaching material repository

The repository to share teaching material among participants of the Dagstuhl Seminar was set up using the mini-MOOC software of the Biolab group of Blaz Zupan. The repository is hosted under `http://books.biolab.si/books/biomedvis`. The overall repository is classified as CC BY-NC-SA 4.0, however, each contributor should also make sure that the copyrights of the shared material are clearly marked.

Last but not least, we discussed a possible future direction or extension of our working group, which is the creation of an interactive, open textbook. This book could encompass the following additional materials and approaches:

- lecture notes for each component of the syllabus,
- short accompanying videos,
- interactive content, including
  - quizzes with progress monitoring and authentication, where needed,
  - text-based answers with AI answer verification, hints, and critique,
  - gamification,
- proposals for group activities during teaching.

References

1 Bach et al., Challenges and Opportunities in Data Visualization Education: A Call to Action, https://doi.org/10.48550/arXiv.2308.07703, to appear in IEEE TVCG (paper at VIS 2023).

2 Bloom, Benjamin S and Krathwohl, David R, Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain. Longmans 1956.

4.3 Beyond the Desktop: Leveraging Immersive Environments for Biomedical Data Analysis – Challenges, Vision, and Guidelines

Michael Krone (Universität Tübingen, DE), Jillian Aurisano (University of Cincinnati, US), Marc Baaden (Laboratoire de Biochimie Théorique – Paris, FR), Nadezhda T. Doncheva (University of Copenhagen, DK), Zeynep Gümüs (Icahn School of Medicine at Mount Sinai – New York, US), Ingrid Hotz (Linköping University, SE), Tobias Isenberg (INRIA Saclay – Orsay, FR), Karsten Klein (Universität Konstanz, DE), Torsten Kuhlen (RWTH Aachen, DE), Trevor Manz (Harvard University – Boston, US), Scooter Morris (University of California – San Francisco, US), Bruno Pinaud (University of Bordeaux, FR), Falk Schreiber (Universität Konstanz, DE), and Anders Ynnerman (Linköping University, SE)

Our group included participants with extensive experience collaborating with domain scientists in designing visualization applications and environments for biomedical applications. Participants also had extensive experience creating and running visualization facilities with diverse display platforms (CAVEs/CAVE2, display walls, DOMEs, touch-tables, VR headsets, AR-capable devices).

This subgroup discussed challenges in immersive analytics and honed in on a specific and compelling challenge that is relevant for many biomedical analysis scenarios: integration of visualizations of abstract data- often represented in 2-dimensions- along with visualizations of spatial data- often represented in 3-dimensional views and potentially within display platforms capable of stereoscopic 3D and immersive presentation and exploration (e.g., see Fig. 6. These immersive platforms include (but are not limited to) VR headsets, CAVEs/CAVE2, stand-alone stereoscopic monitors or display walls, and AR-capable devices, as well as multimodal devices making use of haptic feedback and other diverse interaction modalities. From our experience running visualization facilities and collaborating with domain scientists, we have noticed a benefit for viewing this data in 3D-stereoscopic environments. However, there are challenges that currently limit design and development in this space. These include interaction design that addresses the unique affordances of 2D and immersive environments, coherently linking abstract and spatial data and in supporting realistic workflows. In addition, evidence on the quality and usability of approaches and designs is scarce and not yet organized well in the context of the design space. This hinders comparison between different approaches and to draw insights from existing designs for the development of new designs for differing but potentially related use-cases.

We proposed proceeding from biomedical use-cases which featured a need for tasks that spanned abstract and spatial data at the same time. We intend to characterize a visualization design space to address the common needs and challenges across these use cases. We intend to analyze three different groups of use cases in more detail

- Visual analysis of molecular structures requiring advanced interaction including a variety of 2 dimensional representations as statistical plots or networks.
- Visualization and analysis of Cell models including Metabolic/PPI network
- Visualization of brain imaging data in relation to different diseases/conditions integrating existing knowledge about e.g. brain atlas. Such data frequently includes a large variety of clinical data e.g. questionnaires, and blood samples.

These use cases were selected because they involve abstract and spatial data, diverse users, large scales and high complexity. These focused use-cases will enable us to bring to the surface critical challenges for visualization design and development that integrates abstract and spatial data, within 2D and 3D views.

The concepts we intend to consider include transitional and hybrid interfaces as well as a single unified environment (e.g. to bring the 2D into 3D views such as VR, or in a Dome setup) that combines modalities as design options. In these considerations it is important to factor in specific constraints related to the use case, for instance on readability of the 2D content if text is involved. Our group arrived at a consensus that for this proposed work we will not focus on one platform. Different platforms present trade-offs for users and developers. Rather than explore the design space for one platform, we intend to help designers, developers and users consider when, where and how to use different platforms for different cases.

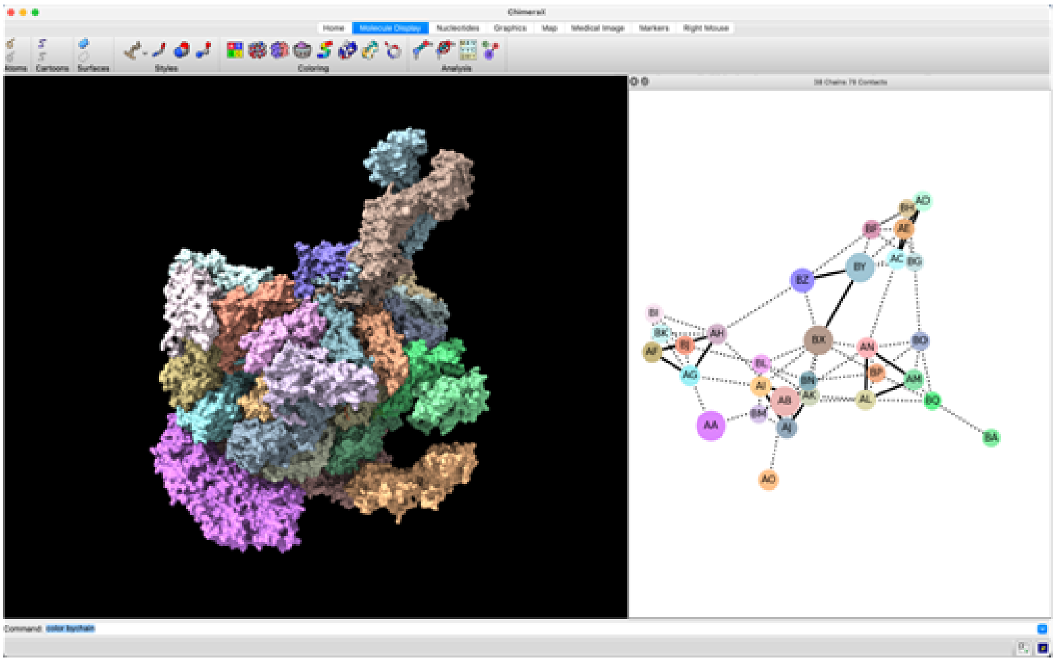


Figure 6 Example of a 2D rendering of a complex molecular structure in relation to a linked interaction network (generated using ChimeraX by Scooter Morris).

We intend to explore design space options for tasks in BioMedVis that span 2D and 3D views. Preliminary design space divisions include where to present 2D content in relation to 3D content. Some options include bringing 2D content into a 3D space- either by overlaying 2D content on the 3D content through visual channels or interactions- or presenting 2D content adjacent to 3D content on a 2D plane. A second option involves juxtaposing 2D views with 3D views, on separate devices and then considering interactions to support integration.

A different design space consideration involves roles for 3D views and 2D view, such as 2D views supporting interactive selections, application of filters, scales and aggregations which modulates what is presented in 3D views.

4.4 Bridging Scales, Data, and Modalities

Michael Krone (Universität Tübingen, DE), Jan Aerts (Amador Bioscience – Hasselt, BE), Jillian Aurisano (University of Cincinnati, US), Marc Baaden (Laboratoire de Biochimie Théorique – Paris, FR), Nadezhda T. Doncheva (University of Copenhagen, DK), Zeynep Gümüs (Icahn School of Medicine at Mount Sinai – New York, US), Thomas Höllt (TU Delft, NL), Ingrid Hotz (Linköping University, SE), Karsten Klein (Universität Konstanz, DE), Torsten Kuhlen (RWTH Aachen, DE), Devin Lange (University of Utah – Salt Lake City, US), Trevor Manz (Harvard University – Boston, US), Scooter Morris (University of California – San Francisco, US), Ramasamy Pathmanaban (Ghent University, BE), Bruno Pinaud (University of Bordeaux, FR), and Falk Schreiber (Universität Konstanz, DE)

The bridging scales group met to discuss the changing nature of biological visualization. Given the Nature Methods [1, 2, 3, 4, 5, 6] series of papers on Visualizing Biological Data in 2010 as a starting point one can observe that the needs of biologists have changed, both as a result of the massive increase in the scale of the data and the increasing need to cross traditional domain boundaries. The discussion started with reviewing the current landscape of needs for visualization in the biomedical domain. This included a brief overview of data types and scale dimensions where complexity increases:

Biomedical Data Types (can vary with condition and time):

- Structures (e.g., protein structure)
- Omics
- Imaging / Fields
- Clinical data
  - Catch-all for many types of data
  - Text
  - Patient data
  - Just call it “metadata”?
- Interactions / Connections
- Time series (A “continuous” signal over time, e.g., heart beat)
- Text

Scales:

- Spatial
- Hierarchical level
- Temporal
- # Items
- # Dimensions
- Heterogeneity (of different data modalities)

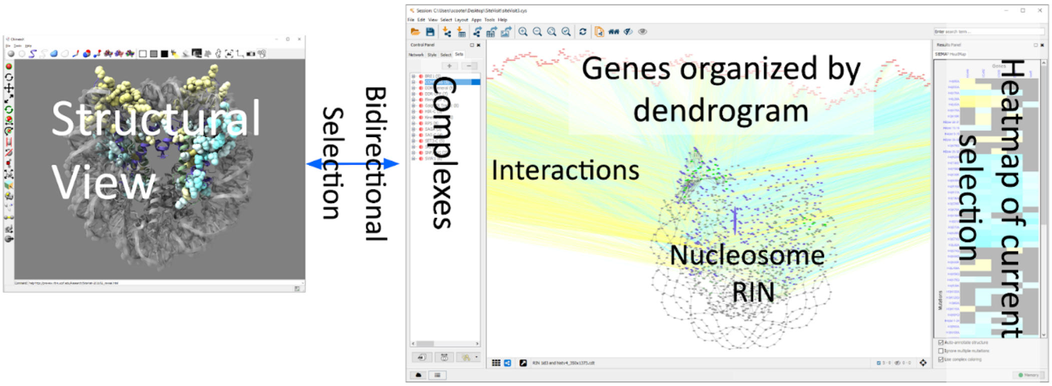


Figure 7 Example pE-MAP.

In order to make the discussions more concrete, we discussed some example use cases, including the interaction between molecular visualization (structures) and networks (e.g. contact networks); Heatmaps (Omics) that vary over time; and pE-Maps (interaction between gene knockouts and point mutations – see image). While discussing the example use cases, the issue of how to approach visualization solutions to these challenges was discussed. This highlighted the issue of developing one-off specialized solutions, potentially multiple times targeting the same challenge, and the tendency for visualization researchers to work in isolation from potential interdisciplinary collaborators. We recognize that having visualization experts is ideal, but there are practical limitations which make it infeasible to make this expertise ubiquitous across biological labs. We discussed mechanisms to spread visualization expertise, such as a centralized visualization service. Additionally we identified our role as a community in educating (e.g., sharing interaction techniques) and designing reusable tools which may be composed by bioinformaticians to tailor custom built applications.

We then focused on the necessity of visualization expertise to create tools that can handle complex data integration, recognizing a general lack of understanding about the importance of visualization in the broader scientific community. It was felt that future publications should emphasize the role of visualization in exploratory analysis and the importance of specialized visualization knowledge during analysis.

This allowed us to focus on what we wanted to produce as an outcome and we decided to focus on a possible update to the Nature Methods series, including a perspectives article and a series of short pieces focusing on practical advice on how challenging examples of complex data from imaging, omics, and computational methods can be visualized and interactively explored. We've written an initial abstract and will be reaching out to the editors of Nature Methods to determine the level of interest.

References

1 O'Donoghue SI, Goodsell DS, Frangakis AS, Jossinet F, Laskowski RA, Nilges M, et al. Visualization of macromolecular structures. Nat Methods. 2010;7: 42.

2 Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. Nat Methods. 2010;7: S56–68.

3 Walter T, Shattuck DW, Baldock R, Bastin ME, Carpenter AE, Duce S, et al. Visualization of image data from cells to organisms. Nat Methods. 2010;7: S26–41.

4 Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. Nat Methods. 2010;7: S16–25.
5 Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T. Visualizing genomes: techniques and challenges. Nat Methods. 2010;7: S5–S15.
6 O'Donoghue SI, Gavin A-C, Gehlenborg N, Goodsell DS, Hériché J-K, Nielsen CB, et al. Visualizing biological data–now and in the future. Nat Methods. 2010;7: S2–S4.

4.5 The Role of Data Visualization in Instigating Behavior Change to Promote Healthy Lifestyles

Cagatay Turkay (University of Warwick – Coventry, GB), Carsten Görg (University of Colorado – Aurora, US), Lynda Hardman (CWI – Amsterdam, NL & Utrecht University, NL), Devin Lange (University of Utah – Salt Lake City, US), Trevor Manz (Harvard University – Boston, US), Georgeta Elisabeta Marai (University of Illinois – Chicago, US), Anders Ynnerman (Linköping University, SE), and Xiaoru Yuan (Peking University, CN)

Theme of the discussion

A common problem is that people may not want to adapt their lifestyle in a way that would benefit their health. They could be unwilling for different reasons, such as an unwillingness to believe the evidence, or may have difficulty understanding it. Perhaps they are convinced but have difficulty maintaining the required lifestyle changes long term.

Our approach is to consider how data visualizations can be used in a dialogue with individuals as a fundamental component in preventive healthcare. We envisage a process where people are first able to express their concerns, then gradually explore data visualizations that allow them to discover and understand how specific lifestyle changes influence health. As a secondary topic, we are interested in how longer term engagement can be promoted through data visualizations tailored to individuals and their context.

Our goal is to produce a position paper to be published in a venue with a broad design audience, such as Transactions on Visualisation in Computer Graphics or Computer Graphics and Applications. The paper will draw inspiration from four main fields:

- Factors that influence behavior change and visualizations that can promote this
- Visual communication that can engage and influence skeptical audiences
- Visual communication in public health applications
- Storytelling using data visualizations

Our contributions are a set of considerations that we deem helpful to designers who are faced with the task of designing an application for a specific group of users. We use the considerations to describe a few examples of existing projects where the goal is to change the mindset and/or behavior of the audience. Based on our discussion we provide recommendations for the considerations we identified.

Visualization Design Considerations

Facilitating behavior change is difficult in any setting and in particular in public health. Visualizations often play a role in these types of communications. But to what extent and how they do so is a difficult question to untangle. We have developed an initial set of considerations when developing visualizations for behavior change within public health. These considerations are not exhaustive but are intended as a starting point for aiding in the development of these visualizations. These considerations are described as if they are independent, however in reality, many are tightly interconnected and cannot be separated.

For example, there are various *storytelling strategies* for encouraging behavior change. Some stories may choose to say directly what the desired behavior change would be and back up that claim with data. Alternatively, data, and a means to explore it, may be provided up front with the hope that readers will draw their own conclusions. We don't claim that there is always a correct strategy, but rather this is a consideration to be taken into account when constructing visualizations. This decision often will depend on other considerations, such as the *communication medium* (static image, interactive application, short-form video), as well as the desired change in readers (get vaccinated or reduce risk of diabetes through diet/exercise).

Ultimately, these considerations should inform visualization design decisions. However, there may not be simple rules that can be applied. Visualization designers may need to review and synthesize all of these considerations when making design decisions (which they already do). However, our goal is to provide some additional structure so that designers can use these to think through these considerations more systematically and can serve to inform and educate new visualizations designers.

Participants

- Jan Aerts
Amador Bioscience – Hasselt, BE
- Jillian Aurisano
University of Cincinnati, US
- Marc Baaden
Laboratoire de Biochimie Théorique – Paris, FR
- Stefan Bruckner
Universität Rostock, DE
- Katja Bühler
VRVis – Wien, AT
- Nadezhda T. Doncheva
University of Copenhagen, DK
- Mennatallah El-Assady
ETH Zürich, CH
- Carsten Görg
University of Colorado – Aurora, US
- Zeynep Gümüs
Icahn School of Medicine at Mount Sinai – New York, US
- Lynda Hardman
CWI – Amsterdam, NL & Utrecht University, NL
- Thomas Höllt
TU Delft, NL
- Ingrid Hotz
Linköping University, SE
- Tobias Isenberg
INRIA Saclay – Orsay, FR
- Helena Jambor
Universitätsklinikum TU Dresden, DE
- Karsten Klein
Universität Konstanz, DE
- Barbora Kozlíková
Masaryk University – Brno, CZ
- Michael Krone
Universität Tübingen, DE
- Torsten Kuhlen
RWTH Aachen, DE
- Devin Lange
University of Utah – Salt Lake City, US
- Trevor Manz
Harvard University – Boston, US
- Georgeta Elisabeta Marai
University of Illinois – Chicago, US
- Scooter Morris
University of California – San Francisco, US
- Kay Katja Nieselt
Universität Tübingen, DE
- Ramasamy Pathmanaban
Ghent University, BE
- Bruno Pinaud
University of Bordeaux, FR
- James Procter
University of Dundee, GB
- Renata Georgia Raidou
TU Wien, AT
- Timo Ropinski
Universität Ulm, DE
- Falk Schreiber
Universität Konstanz, DE
- Thomas Schultz
Universität Bonn, DE
- Matthias Stahl
DER SPIEGEL – Hamburg, DE
- Cagatay Turkay
University of Warwick – Coventry, GB
- Pere-Pau Vazquez
UPC Barcelona Tech, ES
- Anders Ynnerman
Linköping University, SE
- Xiaoru Yuan
Peking University, CN
- Blaz Zupan
University of Ljubljana, SI

Report from Dagstuhl Seminar 23452

Human-AI Interaction for Work

Susanne Boll*[1], Andrew L. Kun*[2], Bastian Pfleging*[3], and Orit Shaer*[4]

1 **Universität Oldenburg, DE.** `susanne.boll@uni-oldenburg.de`
2 **University of New Hampshire – Durham, US.** `andrew.kun@unh.edu`
3 **TU Bergakademie Freiberg, DE.** `bastian.pfleging@informatik.tu-freiberg.de`
4 **Wellesley College, US.** `oshaer@wellesley.edu`

Abstract

Work is changing. Who works, where and when they work, which tools they use, how they collaborate with others, how they are trained, and how work interacts with well-being – all these aspects of work are currently undergoing rapid shifts. A key source of changes in work is the advent of computational tools that utilize artificial intelligence (AI) technologies. AI will increasingly support workers in traditional and non-traditional environments as they perform manual-visual tasks as well as tasks that predominantly require cognitive skills.

Given this emerging landscape for work, the theme of this Dagstuhl Seminar was human-AI interaction for work in both traditional and non-traditional workplaces, and for heterogeneous and diverse teams of remote and on-site workers. We focused on the following research questions:

- How do we allocate tasks between humans and automation in practical settings?
- How can interfaces allow for the appropriate level of human understanding of the roles of human and machine, for the appropriate trust in machines, and how can they reduce incorrect use and confusion?
- How do we support user attention for different tasks, teams, and work environments?
- How can human-automation interaction technology support both work and worker well-being?

At the seminar, we discussed these questions considering their interconnected nature. This focus on interconnectedness of issues was supported by the interdisciplinary group at the Dagstuhl Seminar which was attended by computer scientists/engineers, electrical engineers, human factors engineers, interaction designers, UI/UX designers, and psychologists from industry and academia.

In the following, we report the program, activities, and outcome of our Dagstuhl Seminar 23452 "Human-AI Interaction for Work."

Seminar November 5–10, 2023 – https://www.dagstuhl.de/23452
2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI)
Keywords and phrases future of work, human-ai interaction
Digital Object Identifier 10.4230/DagRep.13.11.20

* Editor / Organizer

Human-AI Interaction for Work, *Dagstuhl Reports*, Vol. 13, Issue 11, pp. 20–71
Editors: Susanne Boll, Andrew L. Kun, Bastian Pfleging, and Orit Shaer
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Susanne Boll
Andrew L. Kun
Bastian Pfleging
Orit Shaer

Broadly, seminar participants collaborated in asking: "What are the knowledge gaps that we collectively face regarding the design, implementation techniques, and evaluation methods and instruments for novel models of human-AI collaboration for work?" We approached this broad question by focusing on the following interrelated specific research questions. Figure 1 provides an overview of the initially scheduled seminar activities. The subsequent sections outline the different activities and the results obtained in these sessions.

Time	Monday	Tuesday	Wednesday	Thursday	Friday
07:30-08:45	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast
9:00	Welcome, Introduction to Dagstuhl, Get to know each other	Unconf / world cafe	Presentations of Human AI for Work Videos	Report - personal, video, table discussion	Finalize report, Presentations (group work) Future work
10:15-10:45	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
10:45-12:15	Joint Activity	Ideation Session Human AI for Work		CHIWork Presentations + Discussion	
12:15	Lunch	Lunch	Lunch	Lunch	Lunch
14:00	Pecha Kucha 2 Panels	Prototyping Human AI for Work	Group photo. Joint afternoon event: Visit of Völklinger Hütte (World Heritage Site)? Joint dinner (self-paid).	Wild predictions about AI for work	Departure
15:30-16:00	Coffee & cake	Coffee & cake		Coffee & cake	
16:00-18:00	Pecha Kucha 2 Panels Wrap-up & Planning	Prototyping Human AI for Work		Wild predictions about AI for work	
18:00	Dinner	Dinner		Dinner	
19:30	Evening activity	Movie night			

Figure 1 Schedule of Dagstuhl Seminar 23452 (November 5–10, 2023).

- **RQ1 How do we allocate tasks between humans and automation?** Automated systems have been around for decades, however today, computer-based automated apps and devices are woven into our professional lives to a greater extent than before. Our dependency on automated systems such as conversational agents, expert systems, vehicles, and drones in daily tasks will likely increase shortly. This will require new forms of human-automation interaction, allowing us to make decisions and collaborate with automation to achieve some goals. A key question in designing this interaction is how to divide tasks between the human and AI. In many practical settings task division is a difficult problem [1, 2] – we explored how we can create guidelines for task division in various work-related contexts.

- **RQ2 How can interfaces allow for the appropriate level of human trust in machines, and reduce incorrect use and confusion?**
  Whenever automation is involved, we need to design user interfaces that support what Lee and See call calibrated trust [3] – a level of trust that is appropriate for the capabilities of the automated system. If the level of trust is not calibrated, human-AI interaction can suffer in two ways. In the case that the human user has too much trust in the AI, they will tend to accept AI suggestions and decisions without a sufficient level of critical reflection, and in some cases, this will lead to accepting bad AI suggestions or decisions. If, on the other hand, the human has too little trust in the AI, they will ignore valuable input from the AI. We explored human-AI interaction designs that allow users to appropriately calibrate their level trust in the AI.
- **RQ3 How do we support user attention?** The broad question of attention is relevant in many work contexts – in mobile environments like an automated vehicle where the user might have to drive some of the time [4, 5], and at the home office, where multiple distractions could compete for the user's attention [6].
- **RQ4 How do we create and leverage new human-automation interaction technology, and support both work and worker wellbeing?**
  How can technologies such as speech interaction, augmented and virtual reality, and tangible interfaces support human-automation interaction? How can we assure that the technologies are used ethically? Furthermore, as Yuval Noah Harari points out in his book "21 lessons for the 21st century" [7], AI might soon become better than we are at many tasks. How can human users best use, collaborate with, and benefit from such super-smart AI?

References

1 Teodorovicz, T., Sadun, R., Kun, A. L., and Shaer, O. (2021). Working from home during COVID19: Evidence from time-use studies. Harvard Business School.

2 Teodorovicz, T., Kun, A. L., Sadun, R., and Shaer, O. (2022). Multitasking while driving: A time use study of commuting knowledge workers to assess current and future uses. International Journal of Human-Computer Studies, 162, 102789.

3 Michal Rinott, Shachar Geiger, Neil Nenner, Ori Topaz, Ayelet Karmon, and Kiersten Blake. 2021. Designing an embodied conversational agent for a learning space. In Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21). Association for Computing Machinery, New York, NY, USA, 1324–1335. https://doi.org/10.1145/3461778.3462108

4 Uwe Gruenefeld, Lars Prädel, Jannike Illing, Tim Stratmann, Sandra Drolshagen, and Max Pfingsthorn. 2020. Mind the ARm: realtime visualization of robot motion intent in head-mounted augmented reality. In Proceedings of Mensch und Computer 2020 (MuC '20). Association for Computing Machinery, New York, NY, USA, 259–266. https://doi.org/10.1145/3404983.3405509

5 Drolshagen S, Pfingsthorn M, Hein A. Context-Aware Robotic Assistive System: Robotic Pointing Gesture-Based Assistance for People with Disabilities in Sheltered Workshops. Robotics. 2023; 12(5):132. https://doi.org/10.3390/robotics12050132

6 Shruti Mahajan, Khulood Alkhudaidi, Rachel Boll, Jeanne Reis, and Erin Solovey. 2022. Role of Technology in Increasing Representation of Deaf Individuals in Future STEM Workplaces. In Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '22). Association for Computing Machinery, New York, NY, USA, Article 16, 1–6. https://doi.org/10.1145/3533406.3533421

7 Puiutta, E., Abdenebaoui, L., and Boll, S. (2023). The Importance of Trust and Acceptance in User-Centred XAI – Practical Implications for a Manufacturing Scenario. CHI 2023 Workshop on Trust and Reliance in AI-Human Teams (TRAIT).

2 Table of Contents

3 Poem about our Dagstuhl Seminar

Prompt to chat gpt: "Write a poem about a seminar at the Dagstuhl castle (which is called Schloss Dagstuhl in German) where human-computer researchers gathered in November 2023 to discuss how to design interactions between humans and artificial intelligence so as to support the future of work. The researchers and practitioners came from all over the world – from host nation Germany, from Europe from the US from India, and from Australia. They discussed opportunities and pitfalls in using artificial intelligence (which is also called AI) in future situations for work. They visited an old steel plant that is now closed and which in the late 1800s was the peak of technological development. Today the abandoned steel plant looks menacing and serves as an example of inhumane working conditions. This visit served as a warning sign for them because perhaps their own ideas will be judged as inhumane in 10" years. How can they avoid this fate? Only through careful scientific endeavor and user-centered research."

```
In Dagstuhl's ancient halls, minds convened,
Scholars, thinkers, in discourse keen.
Human-computer realms unfold,
In Schloss Dagstuhl, tales untold.

From Germany's embrace and Europe's expanse,
To the US, India, and Australia's dance.
They gathered, diverse, a global array,
To shape the future where AI holds sway.

November's chill embraced the air,
As they pondered, minds afire with care.
Future work, a theme profound,
In Dagstuhl's castle, wisdom found.

A steel plant relic, echoes of the past,
Witness to progress, shadows cast.
Once pinnacle in the 1800s' grace,
Now abandoned, a chilling embrace.

Inhuman echoes, harsh and cold,
The steel plant's tale, a story told.
Caution sounded in the whispers there,
As the researchers sought paths rare.

A warning from history's solemn gaze,
That judgments linger through time's haze.
Could their creations, a decade hence,
Be seen as seeds of inhuman pretense?

To avert this fate, they vowed to strive,
Through science's lens and research alive.
User-centered, their guiding light,
To navigate the future, clear and bright.
```

4 Outlook on the Future of Human-AI Interaction: Participant Statements

4.1 Human-centric design of explainable AI systems for increased trust and acceptance in workplace decision making

Larbi Abdenebaoui (OFFIS – Oldenburg, DE)

Our research addresses the need for Explainable AI (XAI) to cultivate trust and acceptability in the field of workplace automation, with a particular focus on industrial workers. As AI methods proliferate, their inevitable integration into various work domains becomes apparent. The planned study investigates the use of AI to assist human workers in decision-making in production scenarios where factors and constraints resist full digital capture.

We aim to develop a batch ordering system that not only suggests solutions but also explains the rationale behind them, thus facilitating worker training to improve decision-making skills. Throughout the integration and interaction process, key questions arise about the nature of human-AI interaction, the role of explanation in generating trust and acceptance, and the ownership of AI agents – whether they belong to employers or employees.

Addressing these issues requires a multidisciplinary approach that combines ethnographic studies, co-design methods, and prototyping. By using these methods synergistically, we aim to unravel the complex dynamics between AI and industrial workers. We argue that a well-designed system that incorporates explanatory elements can play a central role in cultivating trust and acceptance among workers, thereby mitigating concerns about job displacement and fostering harmonious human-AI collaboration in the workplace.

4.2 Creating human-AI work environments that cherish work

Susanne Boll (University of Oldenburg, DE)

Digitalized tools and systems for and around work pervade many of our work environments. In many fields of work, we perceive that this could be a positive contribution to work and make work more effective, more efficient, and create a better output. So what we see is that performance is one of the key factors here. Can we attend to more customers and offer them a better service and sell more products, can we care for more patients at the same time, can we automate manufacturing processes and produce more products of higher quality with less effort of human work time? We can go on with these kind of expectations for the future of our work. We can also observe that such digitization and automation are coming with fear of unemployment. Some years ago economists investigated Automation Angst and the fear that people would lose their jobs but this did not become true rather job descriptions have changed. It could also be shown that digitalization and automation are not killing jobs but job profiles would change and only some of the work will be automated while workers will attend to different tasks in their work life. What is still not attended to in work life is the way on how this changes our work and how this work can still be meaningful. In Human-Computer Interaction we have many methods and instruments to design interactive systems that are efficient, effective, satisfying, and also joyful. But human work also contributes to one's self-determination. What makes me proud about my work, is how to avoid the failures and shortcomings of digital environments that lead to what is often referred to as human errors, what makes us feel valued and that we receive appreciation for our work. But this is not covered yet by our measures yet. In the field of work they are focusing on the value of work for the individual but not necessarily for digitized work processes. At the same time in HCI we look at usability and UX but only start to look at the bigger picture of how we create meaningful digitized workplaces of the future.

4.3 Enhancing Academic Peer Review with Generative AI: Opportunities and Ethical Considerations

Duncan Brumby (University College London, GB)

Facing a surge in submissions and a shortage of expert reviewers, this abstract suggests employing generative AI to refine and expedite the peer-review process. AI has the potential to clarify and temper reviewers' feedback, fostering better dialogue between authors, reviewers, and editors, and enabling editors to consolidate a variety of reviews for efficient decision-making.

The ethical implications and the possibility of over-reliance necessitate a careful evaluation of AI's place in the peer review process. Assessing how reviewers adopt AI and how authors perceive it, especially in light of some authors' adverse reactions to presumed AI-assisted reviews, is crucial. There is an immediate need for focused research to investigate how generative AI is being used in peer review, to identify best practices, and to formulate new editorial policies concerning AI use. Addressing this issue is urgent, yet it holds the promise of mitigating the ongoing crisis in peer review.

This abstract, developed with the support of Chat GPT 4.0, illustrates the practical advantages of AI in academic discourse. But this hand-written note was still penned and carefully checked by me, the human author.

4.4 Towards Inclusive Innovation: Interrogating Power Dynamics in the Future of Work

Marta Cecchinato (University of Northumbria – Newcastle upon Tyne, GB)

As technology reshapes the future of work, it raises pressing concerns around exacerbating inequalities and imbalances of power. Different working domains highlight ways in which work can both become a level playing field, as well as exacerbate some inequalities. For example, platform-based gig work promises new economic opportunities and can offer income opportunities for those who can struggle to find work elsewhere. However, these platforms lack transparency around the algorithms matching workers and clients and create a number of power imbalances. This risks further exploiting already vulnerable groups [1, 2, 3]. In addition, technology has innovated to ease the burden of unpaid caring labor, but not eliminate it; there is also a risk of over-reliance on imperfect, partial technological substitutes for human care [6]. In another example, neurodivergent individuals who make up a considerable amount of the workforce (15 percent in the UK alone), are expected to produce outputs at the same pace and in the same way as neurotypical workers despite differing needs. Despite this, tools are not designed with neurodivergent individuals in mind and AI offers the potential to support these workers in more efficient ways [4]. Moreover, neurodiversity is often studied in younger populations such as children, and there is less of an understanding and a lack of support in adult populations [4, 5]. Across these examples we see technologies transforming work: on one side, this increases access to work and lowers the burden, but on the other side it changes work in deeply unequal ways. This raises critical research questions around how to increase algorithmic transparency, design human-centred automation, and account for diverse needs in technological systems. To achieve equitable progress, we must build future of work systems accounting for diverse users and well-being, not just technical capabilities. This margins-first, human-centred approach is essential to reconcile innovation's promise with concerns around its unequal impacts, as we shape work's future.

References

1. Alvarez De La Vega, J., Cecchinato, M., & Rooksby, J. (2021). "Why lose control?" A Study of Freelancers' Experiences with Gig Economy Platforms.
2. Alvarez De La Vega, J. C., Cecchinato, M. E., & Rooksby, J. (2022,). Design Opportunities for Freelancing Platforms: Online Freelancers' Views on a Worker-Centred Design Fiction. In 2022 Symposium on Human-Computer Interaction for Work (pp. 1-19).
3. Alvarez de la Vega, J. C., Cecchinato, M. E., Rooksby, J., & Newbold, J. (2023). Understanding Platform Mediated Work-Life: A Diary Study with Gig Economy Freelancers. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-32.
4. Campbell, J, Çerçi, S., Cecchinato, M.E. (2023). ADHD and Knowledge Work: Exploring Strategies, Challenges and Opportunities for AI. Designing Technology for Neurodivergent Self-Determination: Challenges and Opportunities workshop, Interact 2023
5. Spiel, K., Hornecker, E., Williams, R. M., & Good, J. (2022, April). ADHD and technology research–investigated by neurodivergent readers. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-21).
6. Wang, J., O'Kane, A. A., Newhouse, N., Sethu-Jones, G. R., & De Barbaro, K. (2017). Quantified baby: parenting and the use of a baby wearable in the wild. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1-19.

4.5 Can Sensing Technologies Augment Remote and Hybrid Work Without Crossing Ethical Lines?

Marios Constantinides (Nokia Bell Labs – Cambridge, GB)

The global shift to remote and hybrid work, accelerated by the COVID-19 pandemic, has challenged and disrupted traditional work norms; from the well-known eight-hour workday within the confines of the office to the salient boundaries between work and personal life [1]. This transformation prompts profound inquiries into how to support and augment workers' experience through tools that not only improve their productivity but also enhance their emotional and psychological well-being. A case in point is reimagining work meetings through the lens of mobile and wearable sensing technologies. These technologies allow us to tap into workers' emotional states [2], assess their environment [3], and understand communication patterns that might go unnoticed [4]. At the same time, these technologies are not without ethical considerations. As we move towards a future dominated by big data and powerful AI algorithms, a host of new questions arises regarding the psychological impact of workplace surveillance, data governance, and the compliance of these technologies with ethical and moral concerns [5].

References

1 Rudnicka, A., Newbold, J., Cook, D., Cecchinato, M., Gould, S. & Cox, A. Eworklife: Developing effective strategies for remote working during the COVID-19 pandemic. *Eworklife: Developing Effective Strategies For Remote Working During The COVID-19 Pandemic.* (2020)
2 Park, S., Constantinides, M., Aiello, L., Quercia, D. & Van Gent, P. Wellbeat: A framework for tracking daily well-being using smartwatches. *IEEE Internet Computing.* **24**, 10-17 (2020)
3 Constantinides, M., Šćepanović, S., Quercia, D., Li, H., Sassi, U. & Eggleston, M. ComFeel: Productivity is a matter of the senses too. *Proceedings Of The ACM On Interactive, Mobile, Wearable And Ubiquitous Technologies.* **4**, 1-21 (2020)
4 Choi, J., Constantinides, M., Joglekar, S. & Quercia, D. KAIROS: Talking heads and moving bodies for successful meetings. *Proceedings Of The 22nd International Workshop On Mobile Computing Systems And Applications.* pp. 30-36 (2021)
5 Constantinides, M. & Quercia, D. Good Intentions, Bad Inventions: How Employees Judge Pervasive Technologies in the Workplace. *IEEE Pervasive Computing.* **22**, 69-76 (2023)

4.6 Academic AI Interactions

Anna Cox (University College London, GB)

Main reference Joseph W. Newbold, Anna Rudnicka, David Cook, Marta E. Cecchinato, Sandy J. J. Gould, Anna L. Cox: "The new normals of work: a framework for understanding responses to disruptions created by new futures of work", Hum. Comput. Interact., Vol. 37(6), pp. 508–531, 2022.
URL https://doi.org//10.1080/07370024.2021.1982391

Within higher education, the issue of overwhelming workloads and associated stress is a significant concern. Notably, in 2004, the majority of university professors in the U.K. (62 percent) reported working more than 48 hours weekly (Kinman et al 2006). This upward trend persisted, as indicated by a 2016 survey (UCU 2016), which documented that over 12,000 academic staff from various disciplines averaged 50.9 working hours per week. Such figures represent 139 percent of the nominal working week designated for academic staff at UCL, surpassing the maximum allowed by The Working Time Regulations as set by the UK Government.

A plethora of elements contribute to the unsustainable workloads shouldered by academics. Expanding role responsibilities, fluctuating staff numbers, growing student populations, and heightened student expectations have all played a part (UCU 2016). Tight (2010) posits that a considerable portion of the perceived workload increase stems from a surge in administrative tasks, while Miller (2019) contends that, contrary to the widely claimed 40:40:20 division of teaching, research, and administrative duties, a more realistic distribution is 40:25:35. Consequently, fulfilling research output expectations may be unfeasible within the allotted time.

More recently, the tempo and pressure of academic duties have intensified further (Ylijoki 2013). A significant contributing factor is the surge in the use of digital communication technologies, which has been associated with increased feelings of overload and diminished psychological disconnection from work, a crucial aspect of recuperation and burnout avoidance (Sandoval-Reyes et al 2019). Additionally, the uptick in digital tool utilization, a response to the COVID-19 pandemic's push towards remote and hybrid work environments, is likely intensifying the perceived workload even more (Raghavan et al 2021).

Artificial Intelligence tools hold the promise of transforming the workplace by expediting tasks and conserving time creating yet another new normal of work for academics. Despite this, the focus within universities has been more on how Large Language Models (LLMs) will be integrated by students, prompting measures to govern the appropriate use of such innovations (Shearing and McCallum 2023). Although efforts have been made to acquaint academics with the potential benefits of LLMs in their professional activities (Dianti & Laudari 2023a, 2023b, 2023c), research is lacking on the degree of adoption of these tools within academic circles and their impact on work practices. Previous research suggests that workers vary in terms of the strategies they adopt and the journeys they go through when adapting to disruptions to work (Newbold et al 2022).

```
**Academic AI Interactions:
a poem written in collaboration with ChatGPT 4**

In lecture halls and campus walks, The clock ticks on, the professor balks.
Heavy is the head that grades the test, In academia, they've scarce time to rest.
A surging tide of tasks and student needs, On every hour, the academic feeds.
From admin desks to virtual space, The workload mounts a relentless pace.
With emails, pings, the tech's embrace, The prof's measured steps turn to a race.
AI promises a future bright, To lift the load, to ease the plight.
Yet in the groves of Academe, It's more a dream, less a theme.
Where students learn the LLM's sway, Professors yearn for light of day.
To teach, to write, to find reprieve, In AI's web, they must believe.
```

References

1 Dianati, S. & Laudari, S. (2023a) ChatGPT and generative AI: 25 applications in teaching and assessment https://www.timeshighereducation.com/campus/chatgpt-and-generative-ai-25-applications-teaching-and-assessment

2 Dianati, S. & Laudari, S. (2023b) ChatGPT and generative AI: 25 applications to support administrative tasks https://www.timeshighereducation.com/campus/chatgpt-and-generative-ai-25-applications-support-administrative-tasks

3 Dianati, S. & Laudari, S. (2023c) ChatGPT and generative AI: 25 applications to support student engagement https://www.timeshighereducation.com/campus/chatgpt-and-generative-ai-25-applications-support-student-engagement?utm_source=newsletter&utm_medium=email&utm_campaign=campus-weekly&spMailingID=27544421&spUserID=MTAxNzcxMzU4MDQ4NwS2&spJobID=2331588124&spReportId=MjMzMTU4ODEyNAS2

4 Kinman, G., Jones, F., & Kinman, R. (2006). The well-being of the UK academy, 1998–2004. Quality in higher education, 12(1), 15-27.

5 UCU (2016). Workload is an education issue https://www.ucu.org.uk/media/8195/Workload-is-an-education-issue-UCU-workload-survey-report-2016/pdf/ucu_workloadsurvey_fullreport_jun16.pdf.

6 Miller, J. (2019). Where does the time go? An academic workload case study at an Australian university. Journal of Higher Education policy and management, 41(6), 633-645.

7 Newbold, J. W., Rudnicka, A., Cook, D., Cecchinato, M. E., Gould, S. J., & Cox, A. L. (2022). The new normals of work: a framework for understanding responses to disruptions created by new futures of work. Human–Computer Interaction, 37(6), 508-531.
8 Raghavan, A., Demircioglu, M. A., & Orazgaliyev, S. (2021). COVID-19 and the new normal of organizations and employees: An overview. Sustainability, 13(21), 11942.
9 Sandoval-Reyes, J., Acosta-Prado, J. C., & Sanchís-Pedregosa, C. (2019). Relationship amongst technology use, work overload, and psychological detachment from work. International Journal of Environmental Research and Public Health, 16(23), 4602
10 Shearing, H.and McCallum, S. (2023) ChatGPT: Can students pass using AI tools at university? `https://www.bbc.com/news/education-65316283`
11 Tight, M. (2010). Are academic workloads increasing? The post-war survey evidence in the UK. Higher Education Quarterly, 64(2), 200-215
12 UK Gov `https://www.gov.uk/maximum-weekly-working-hoursaccessed2023`
13 Ylijoki, O. H. (2013). Boundary-work between work and life in the high-speed university. Studies in Higher Education, 38(2), 242-255.
14 Sandoval-Reyes, J., Acosta-Prado, J. C., & Sanchís-Pedregosa, C. (2019). Relationship amongst technology use, work overload, and psychological detachment from work. International Journal of Environmental Research and Public Health, 16(23), 4602

4.7 Transforming Healthcare: The Impacts of Recent AI Advances on Healthcare Professionals and Stakeholders

Mohit Jain (Microsoft Research India – Bangalore, IN)

The rapid progression of AI technologies promises to revolutionize the landscape of healthcare delivery in the near future. We explore the multifaceted disruption AI is poised to bring to the healthcare ecosystem, affecting a wide array of stakeholders, including doctors, nurses, patients, caregivers, and community health workers.

Physicians would leverage AI as a powerful tool for accurate and efficient diagnosis, potentially enhancing their diagnostic capabilities and improving patient outcomes. Patients and caregivers stand to benefit from AI-driven platforms that provide reliable and tailored information regarding symptoms and treatment options, facilitating informed decision-making and empowering individuals to take an active role in their healthcare journey. Community health workers are anticipated to utilize AI to streamline their day-to-day activities, optimizing resource allocation and improving overall operational efficiency. Moreover, AI-driven solutions offer nurses novel opportunities for immersive and personalized training experiences, potentially revolutionizing their professional development.

As AI continues to advance, it is imperative for healthcare professionals and stakeholders to adapt and harness the potential benefits of these technologies. However, it is also crucial to address ethical, regulatory, and privacy concerns to ensure that the integration of AI into healthcare remains responsible and patient centered.

4.8 Designing Human-AI Collaboration for Meaningful Work

Shadan Sadeghian (University of Siegen, DE)

Work is an important part of our lives. It is not only a way to earn a living, but a crucial source for experiencing meaningfulness in everyday life. Meaning in work can stem from various sources: being successful at the job, mastering interesting challenges, having good relationships with colleagues, and feeling proud to contribute and to be a crucial part of an organization. Work has always been mediated by technology, from early hand tools to the now ubiquitous computer. However, these tools were largely passive extensions of the body and the mind of a working person, an extension of the self.

The progress in adaptive automation, and the rise of robots and in the widest sense artificial intelligence-based systems will change the way we work. The computational artifacts become counterparts with their own (limited) agency, unpredictability and opacity. Consequently, this will impact the perception of work meaningfulness and job satisfaction on both individual and societal levels. Interestingly, most research on designing AI-based technologies focused on performance-related aspects, such as the efficiency and effectiveness of human-AI systems. Furthermore, due to the limitations of the existing AI-based technologies, these studies mainly address the interaction with blue collar (manual) workers. However, progress in AI will also affect work practices of people entitled as white collar (knowledge), pink collar (service provider), and even no collar (artists) workers.

In this Dagstuhl Seminar, we addressed and discussed these gaps raising the questions of how can the design of future AI-based technology maintain or even enhance job meaningfulness in different domains of work, and how can we ensure meaningfulness besides aiming for effectiveness and efficiency.

4.9 Future of Work in Mobility Domains

Christian P. Janssen (Utrecht University, NL)

Main reference Christian P. Janssen, Stella F. Donker, Duncan P. Brumby, Andrew L. Kun: "History and future of human-automation interaction", Int. J. Hum. Comput. Stud., Vol. 131, pp. 99–107, 2019.
URL https://doi.org//10.1016/J.IJHCS.2019.05.006

Automation and AI is changing various aspects of our lives [3] and in our work [2]. This also holds for the field of mobility in various ways. AI and automation can change why we move and how we move, but also how we consider mobility at large and how it impacts other facets of our lives.

I am particularly thinking of three types of impact: 1. Automated vehicles might have the potential to take over more of the driving tasks for humans. When they reach a sufficient level of maturity, that might give the human some time to do other tasks, which can turn the car into a "mobile office" [1]. But how should this be designed such that safety is not compromised? 2. Public transportation is dependent on scheduling by others. This currently involves quite a bit of planning, but AI can help with these scheduling issues both on the planner's side ("how to solve planning problem X?") and on the consumer's side ("given my trip, what is the best transportation mode?"). There are technical questions about how to best schedule this efficiently, but also on the human side such as how to handle uncertainty in schedules. This can have an impact on how you schedule your work and the degree to which you are willing to live further from home. 3. In parallel, people might also reconsider whether they want to work at their homes and avoid transportation. This has implications of how transportation at large is organized (e.g., how many roads and trains are needed?)

Such situations not only require the study of technology but also the careful study of behavior and prediction of human behavior (see also seminar 22102 [4] and [5]).

References

1 Chuang, L. L., Donker, S. F., Kun, A. L., & Janssen, C. P. . (2018) Workshop on the mobile office. In Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 10-16).

2 CHIWORK Collective, Karusala, N., Ch, N. A. N., Tosca, D., Ansah, A. A., Brulé, E., ... & Kun, A. L. (2022). Human-Computer Interaction and the Future of Work. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1-3).

3 Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019). History and future of human-automation interaction. International journal of human-computer studies, 131, 99-107.

4 Janssen, C. P., Baumann, M., Oulasvirta, A., Iqbal, S. T., & Heinrich, L. (2022). Computational Models of Human-Automated Vehicle Interaction (Dagstuhl Seminar 22102). In Dagstuhl Reports (Vol. 12, No. 3). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

5 Janssen, C. P., Baumann, M., & Oulasvirta, A. (2024). Computational Models of Cognition for Human-Automated Vehicle Interaction: State-of-the-art and Future Directions. International Journal of Human-Computer Studies, 103230.

4.10 AI and the Future of Care Work

Naveena Karusala (Harvard University – Allston, US)

We are in the midst of a global care crisis. Chronic underinvestment in care economies, such as healthcare and social work, has led to an inability to meet growing care needs, especially in marginalized communities. In response, governments and nonprofits have turned to artificial intelligence (AI) and data-driven tools to help use limited resources for care more efficiently and at scale. For these tools to have the desired impact, they need to support the agency and labor of care workers and the communities they serve. There is also significant potential to investigate the value of these tools beyond efficiency and scale, by centering the vision of communities most impacted by the care crisis. One essential direction of future of work research in Human-Computer Interaction will be to investigate how AI and data-driven tools can address inequities in care economies and serve the priorities of marginalized communities. It is especially important to understand and address these issues from a global perspective, in order to resist hierarchies in policy development, and enable models for greater agency and ownership over technology and data globally.

4.11 Human-AI Collaboration and the Future of Work

Andrew L. Kun (University of New Hampshire, Durham, NH, US)

Joint work of Andrew L. Kun, Orit Shaer

The system of 9-5 jobs in the office for 5 days a week is being replaced by a flexible and distributed system of work [1, 2, 3]. Workers now often work from home, and we can expect that this trend will continue. Thus, our workplaces will become more heterogeneous than today: we can expect distributed teams, collaborating both synchronously and asynchronously.

However, such collaborations can be hard. It is easier to communicate with a co-present colleague than to exchange messages with one or more remote colleagues (e.g. [4, 5]. We can expect that AI tools will help us support these new and growing heterogeneous teams as the collaborate in creative tasks. AI can help guide team interactions, acting as a supportive team members, as a coach, and as a source of positive nudges that affect individual actions, with the goal of improving collective creativity of the team, as well as worker well-being (cf. [6]).

How can we do this? One part of the answer lies in creating the appropriate human-computer interaction. This is what our seminar discussed and where we hope to make a lasting contribution in the coming years.

References

1 Mark, G., Kun, A. L., Rintel, S., and Sellen, A. (2022). Introduction to this special issue: the future of remote work: responses to the pandemic. Human–Computer Interaction, 37(5), 397-403.

2 Kun, A. L., Shaer, O., and Iqbal, S. (2021). The future of work: COVID-19 and Beyond. IEEE Pervasive Computing, 20(4), 7-8.

3 Kun, A. L., Sadun, R., Shaer, O., and Teodorovitz, T. (2020). Where did the commute time go?. Harvard business review.

4 Yang, F., Heeman, P. A., and Kun, A. L. (2011). An investigation of interruptions and resumptions in multi-tasking dialogues. Computational Linguistics, 37(1), 75-104.

5 Ansah, A.A., Xing, Y., Kamaraj, A.V., Tosca, D., Boyle, L., Iqbal, S., Kun, A.L., Lee, J.D., Pahud, M. and Shaer, O., (2022). " I need to respond to this"–Contributions to group creativity in remote meetings with distractions. In Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work (pp. 1-12).

6 Ch, N.A.N., Ansah, A.A., Katrahmani, A., Burmeister, J., Kun, A.L., Mills, C., Shaer, O. and Lee, J.D. (2023). Virtual nature experiences and mindfulness practices while working from home during COVID-19: Effects on stress, focus, and creativity. International Journal of Human-Computer Studies, 171, 102982.

4.12 Human-AI Collaboration to Shape the Future

Sven Mayer (LMU München, DE)

The rise of Artificial Intelligence in all parts of life will profoundly impact work. While this will bring changes to every workplace, we still need to understand how the future workplace could look like. We should be in power to design the future we want – thus, shaping the future is now up to us. But how this is possible is also unclear. I believe that human-AI collaboration is the major aspect of ensuring a net positive outcome and, as such, an excellent tool to shape the future workplace. Designing the collaboration allows us to supervise how the AI impacts our work, consequently allowing us to stay in control but outsource tasks to elevate productivity. Thus, the objective is to empower humans to surpass their performance through seamless interaction with AI.

4.13 If it does not work, it is not usable. If it is not usable, it does not work. How to make interactive systems embedding AI technologies, usable?

Phillippe Palanque (Paul Sabatier University – Toulouse, FR)

Automation has been studied for years (not to say decades) and the development of AI technologies is only a step and a mean to further foster the development of automation. Key aspects of usable automation are transparency (how to present the current state of automation), predictability (how to present future states of automation), and controllability (how to start, stop, tune, parameterize ... automations). These required properties, to make automation usable, impose methods, techniques and tools to guarantee them and, more globally, to guarantee their presence in the entire interactive system. Beyond design aspects of interaction with interactive systems (which may embed AI technologies), implementation aspects (including architectures) have to be considered. This means that research should be carried on that matters to support each and every phase in the development process. Unfortunately, so far, limited research is carried out on that matter where the focus is more on prototypes and user studies.

4.14 Designing meaningful collaborations between users and AI-enabled systems

Bastian Pfleging (TU Bergakademie Freiberg, DE)

With the rise of AI, we expect that the way we use computers will change massively wherever they are used. We assume that this will affect almost any workplace and the way we work. While a lot of advances happened at the technological side, the question is still how we should design systems, interfaces, and processes around AI-enable systems.

A special focus should be on the human-centered design of such systems to ensure that systems are built according to their users' needs and considering the various (ethical, societal, ...) open questions. How this could look like is still unclear and it now our task to wisely fill this gap to ensure that work stays or becomes meaningful. My assumption is that one key to success is to focus on the collaboration between humans and AI and ensuring that the user is in control.

In our seminar, we looked at various aspects and challenges regarding the use of AI for work and engaged in many discussions. I am curious to see how we can take these results and share them with the different fields to positively shape the future of work.

4.15 Designing for Human-AI Synchronization

Michal Rinott (SHENKAR – Engineering. Design. Art – Ramat-Gan, IL)

This is a new community for me, and I am excited about the combination of people with technological and psychological orientations. My own background is cognitive psychology→ UX→ interaction design→ HCI so I feel a strong relevance, as well as the potential of adding my interaction design perspective.

I am interested in the relationship between humans and AIs: interaction paradigms, metaphors, mental models and physical designs, when relevant.

One relevant project of mine is "An embodied conversational assistant for a learning space" presented at DIS 21. We envisioned an AI team member for our own Interaction Design lab, and worked with stakeholders to define interactions that would enhance students' experience. The embodiment and relational aspects of the agent/robot were given a lot of attention: the robot – with its deep knowledge of the lab – helps students with to find components and learn about past projects. In return, it needs their help to move around! A mutually assistive relationship. We envision the design of such AI entities fine-tuned to other work environments, with the attention to the design of relational and character that is coherent with the specific attributes of the work culture and space.

My current research looks at Interpersonal Motor Synchronization through technology (i.e., moving together with another person). Psych studies show that when people are close, they tend to synchronize, but also that when people are brought into synchrony – the become closer, and perform better together! I am designing for interpersonal sync both in co-present and remote situations.

Can synchronizing with AI make us feel closer, more connected, and more effective? I plan to explore different forms of human-AI synchronization and would be happy to collaborate.

4.16 Integrating AI with Care

Neha Kumar (Georgia Institute of Technology Atlanta, GA, USA)

As we navigate cultural and geopolitical contexts of climate crises, war, pandemics, growing inequities, while moving forward on our tech commitments re: AI, where really are we headed? My work started with looking at care infrastructures more traditionally defined, e.g. with personal/public/global health contexts. But then care work and worker-centered perspectives became an added focus, even outside of the health and well-being domains. Also, thinking about the methods we use, and how we integrate care into these, and finally, planetary care/care towards our higher-level, epistemological commitments – this is where our lab's focus has been in recent years. In each of these contexts, we look at the role

AI-based systems are currently playing – to impact communities and their survival, to impact design and the data/research that fuels design, to impact the future of care work in research and practice. It is critical that we bring diverse, disciplinary perspectives to these questions, responding to the need of the hour.”

4.17 Human-AI Interaction for Scientific Discovery

Aaron Quigley (CSRIO | UNSW School of Computer Science and Engineering, Australia)

The scientific discovery process involves numerous stages which can be assisted by human and artificial intelligence interaction. Today, the stages of sketching, prototyping, or even brainstorming in ideation can be assisted with natural language, processing analysis, and foundation models. The process of scientific discovery itself is being accelerated and supported with AI to identify patterns and trends that humans may overlook. Other approaches are analysing research papers data sets and scientific workflows to determine new hypothesis and research questions. These are afforded to researchers to supplement and help inform their existing experimental design process. Areas, such as drug discovery with AI supported analysis of molecular structures, biological pathways, and clinical data are being employed to identify drug candidates. While robotic support for in Lab experimentation is growing in day to day use. Finally, the exploitation of research outcomes can be supported with AI to analyse market trends, potential industry, partners, market demand, and process automation in commercialisation. In practice, AI will become part of the innovative tools and techniques scientists employ in their future of work to augment human capabilities and accelerate research across various disciplines.

4.18 The Bright Future of Work

Stefan Schneegass (University of Duisburg-Essen, DE)

AI has the potential to revolutionize work across multiple domains. However, there is still a big gap between the theoretical and practical potential shown by research and big tech companies alike, and its meaningful application and integration in the context of work. The core challenge here will not be the technical development, but rather achieving a meaningful integration. I do not advocate for an explicit use but rather for a pervasive integration. In the future, users will not always be aware of the AIs they use, similar to how computing devices are used today. In every system (from desktop to embedded systems), some sort of AI will be integrated to seamlessly assist users in completing their tasks. This currently starts with summarizing emails and other texts but will extend across all domains. Many everyday tasks will disappear, taken over by AI. So, will AI take over our jobs?

AI might render some jobs unnecessary, but only a few. I believe that AI has the potential to empower individuals to focus on the essentials of their jobs, provided it is well-integrated. This will not significantly reduce the number of jobs but will allow for higher productivity and a greater focus on the essential aspects of work in the end. Many administrative tasks, however, that are currently performed by humans could be taken over by AI, possibly even without the need for human intervention. A fundamental societal question that will arise in the next 10+ years is whether AI still requires human intervention for some of these tasks. This will be answered in my abstract for the Dagstuhl Seminar in 2033.

4.19 Investigating the use of LLMs in group ideation

Orit Shaer (Wellesley College, US)

Joint work of Orit Shaer, Andrew L. Kun

The growing availability of generative AI technologies such as large language models (LLMs) has significant implications for creative work. My research explores the impact of generative AI technology on the creativity of teams. Specifically, I am interested in the work of knowledge workers who collaborate to solve problems such as designers, scientists, and managers. We study twofold aspects of integrating LLMs into the creative process – the divergence stage of idea generation, and the convergence stage of evaluation and selection of ideas and solutions. There are several open questions that we investigate:

- How to integrate LLMs into the idea generation phase?
- Does the use of LLMs in Brainwriting enhance the ideation process and its outcome?
- How can LLMs support idea evaluation?
- How to train novice knowledge workers to engage effectively and critically with LLMs in the creative process?

More broadly, I am interested in how AI could be integrated into the workplace, mobile environments, and into work processes in a way that empowers knowledge workers and increases dignity, productivity, creativity, and wellbeing. In the seminar, we explored issues related to human-AI work-related collaboration including bias, autonomy, control, responsibility, and accountability. Beyond the exploration of research questions, the seminar has provided insight into new collaborative research methods such as speculative design, scenario planning, and rapid prototyping. I am most grateful for the deep conversations and exchange of ideas. I look forward to exploring new collaborations.

4.20 Human-AI Collaboration in Complex Environments: Advancing Interaction Modalities

Erin T. Solovey (Worcester Polytechnic Institute, US)

We are still facing the classic problem in the field of human-computer interaction where technology is advancing while human capabilities are staying relatively stable. Ultimately, advanced AI capabilities are only valuable if they support user tasks, contexts and capabilities of the humans and teams using them. When humans work closely together, they can pick up subtle cues from their team members and adapt their behavior appropriately. Humans working closely with AI systems may also give off cues, but the algorithms and agents cannot detect these signals and therefore cannot change behavior. Similarly, in distributed teams, where the humans are not co-located, these signals are also lost, leading to reductions in team performance.

Emerging research on novel interaction modalities (e.g. brain-computer interfaces, augmented reality, touch-sensitive fabrics, etc.) could be utilized for more supportive multimodal human-AI communication. AI could adapt the modality of the intervention, alert or communication based on the changing situational context. However, these are not fully taken advantage of today. A key area of interest is in heterogeneous, distributed multi-human teams collaborating with AI-enhanced agents or robots in complex environments. These teams must navigate critical tasks, sometimes in high-risk domains, where human members collaborate with AI systems in decision-making under conditions

4.21 The Future of Industry and Craft Work in Times of AI

Tim Stratmann (OFFIS – Oldenburg, DE)

I work in the field of Human-Robot-Collaboration. During my research projects, I found that the essential component that enables efficient and meaningful collaboration between robots and humans is the understanding of intentions and activities, as well as the state of the environment. The way to get this kind of understanding in the current state of the art is through AI technologies. Anticipating the future, Human-Robot-Collaboration will become a collaboration between Humans and AIs with a physical embodiment, e.g. a robotic system.

I think, the future of blue-collar work in industry and craft will be shaped by the collaboration of human workers and physically embodied AIs. During one of our seminar sessions, we envisioned the workday of an electrician in the year 2058 including exactly this kind of embodied AI in the form of an AI toolbox for the electrician. This Speculative Design session sparked a lot of interesting ideas and abstractions that I like to explore in the future. Who is responsible for errors the Human AI team performs? Who is in control? How transparent should AI decisions be? Should safety-related decisions always be overruled by the AI?

4.22 From Human-Human Collaboration to Human-AI Collaboration

Dakuo Wang (Northeastern University, U.S.)

AI is everywhere. But we have seen so many successful AI algorithms on research papers, yet there are only a limited number of successful AI systems in the real world. Why do AI systems often fail in the real world deployment, and how can we design and develop a successful AI system? This is a fundamental question that human-centered AI researchers and practitioners dedicate to find an answer. In my work, I propose the **Human-AI Collaboration** framework. It is defined as an ultimate design goal: the AI system should be designed to collaborate with instead of competing against human. It also has various measurable metrics: instead of focusing only on the AI algorithm's performance, or only on the human's perceived explainability, we should measure the success of the human-AI collaboration via the metrics that human are already using in their existing workflow, such as work's productivity, efficiency, and quality of the deliverable product. Simply put, the human-AI collaboration team's performance should be measured by the human-human collaboration workflow's original performance metrics.

Human-AI Collaboration is more than an ultimate design goal or a set of evaluation metrics; it is also a theoretical framework that can guide us move forward with the human-centered AI research. We know little of how human works with AI systems (especiall LLM-based AI systems), but we know a lot about how human works with other humans from the existing organizational behavior and cognitive science literature. More relevantly, within the Computer Science and Human-Computer Interaciton domain, we have the Computer-Supported Cooperative Work (CSCW) subfield with decades-long research exploring human-human collaboration and designing CS system to support such collaborations. For example, Paul Dourish and Victoria Bellotti's classic Collaboration Awareness theory and the Olsons' Remote Collaboration framework both can provide systematic and clear guidelines on how to approach the human-AI collaboration design. One could argue that explanable AI (XAI) is merely a case of human-human collaboration awareness that each collaboration partner should be aware of the other's current work status and the next step plan. We can find more analogies between the human-human team collaboration and the human-AI team collaboration.

In summary, I believe that the goal of human-centered AI research should prioritize the successful transformation of research deliverables into real-world AI applications. In order to achieve such a goal, human-AI collaboration is one promising design and research paradigm to move forward.

4.23 The Neuro-Future of Work

Max L. Wilson (University of Nottingham, GB)

Technology is striving to assess our cognitive data [1]: are we focusing? are we tired? are we stressed? An important future of work, with examples that already exist [2], is when employers hold this data over employees. How well does AI interpret the data signals from employees into inferred cognitive activity? What are our bosses assuming about our cognitive activity and expected ways of working? What is AI inferring about individuals and workforces, and their productivity? Beyond this starting concern, there are two key issues I consider to be important about the future of work with AI. 1) We should consider these AIs as constantly evolving, with their own histories and biases, rather than as a static system. Where my concerns above raise issues with how people will be assessed by systems and what employees will want to know about how the systems view their employment, the fact that it will continuously evolve creates a temporal requirement for how employers and employees will interact with the system. 2) It will further be important to understand how AI systems will require new skills from people, but of more interest to me is to see how AI will create new forms of the so-called "google maps" effect [3]: changing, if not elevating, the skills people need in everyday life. Beyond how job descriptions will change, this perspective implies that people will "change into a faster lane", focusing on things that AI enables us to do. Our future skill-sets will change, affecting the demands on our cognitive activity, and the assumptions we should hold about them. Where all forms of work are considered to be becoming more cognitive [4], our abilities and skills will dramatically shift in the background.

In summary, I consider that new AI systems that will make assumptions about us our cognitive activity are arriving a time where our cognitive activity will take a dramatic shift with the change of living with AI. This will be a turbulent time in the future of work, with many ethical and social consequences.

References

1 Wilson, M. L., Midha, S., Maior, H. A., Cox, A. L., Chuang, L. L., & Urquhart, L. D. (2022, April). SIG: Moving from Brain-Computer Interfaces to Personal Cognitive Informatics. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1-4).

2 `https://www.vice.com/en/article/8xkymg/china-brain-wave-hats-helmets-productivity`

3 `https://medium.com/mind-cafe/is-google-maps-making-us-more-lost-than-ever-34997b035584`

4 Sharples, S. (2019). Workload II: A future paradigm for analysis and measurement. In Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VII: Ergonomics in Design, Design for All, Activity Theories for Work Analysis and Design, Affective Design 20 (pp. 489-498). Springer International Publishing.

4.24 Navigating the Role of AI as a Mediator in Human Interactions

Naomi Yamashita (NTT – Kyoto, JP)

The growing integration of AI into human activities brings about various advantages, including enhanced task efficiency and emotional support. However, as AI's role expands to mediate interactions between humans, it also introduces potential challenges. For instance, within the context of global collaboration, AI can assist non-native speakers in improving their language skills. Nonetheless, this may inadvertently lead to misjudgments regarding individuals' actual language proficiency, potentially resulting in misunderstandings – especially in situations where AI usage is not feasible or practical. Similarly, in the field of mental healthcare, the utilization of AI in mediating conversations may pose a risk to trust and privacy. There's a concern that sensitive information shared with AI might be disclosed to others in a manner or language that individuals do not wish for. Looking ahead, when AI serves as an intermediary in human-to-human interactions, it will require more than simply transmitting information received from AI. We must carefully consider the types of interactions necessary to foster seamless connections between individuals while proactively addressing these potential issues.

5 Human and AI as a Team: Peanut Butter and Jelly Session

In this session, we aimed to understand how an AI and a human could work together to create something as easy as a peanut butter and jelly sandwich (PB&J). We defined three tasks which were taken over by three teams of participants. In Team 1, two team members independently used two different LLMs to generate step-by-step instructions for making a

PB&J sandwich. Two other team members then independently followed the instructions exactly as it was written to prepare the sandwich (see Section 5.1). In Team 2, two team members independently used a text-to-image model to generate visual instructions for making a PB&J sandwich (see Section 5.2). Two other team members independently follow exactly the instructions to prepare the sandwich. In Team 3, two team members wrote text-based narrative instructions for a robot to make a PB&J sandwich (see Section 5.3). The robot consists of a camera and two 6DOF robotic arms. Two other team members pretended to be robots and independently followed exactly the instructions as they were written to prepare the sandwich. The whole exercise was very interesting as it revealed how much knowledge and experience is on the human side and how much effort it took to communicate the tasks and the steps to and with the AI. It was also fun!

5.1 Team 1: Text-based instructions by two LLMs

Two team members independently used ChatGPT 3.5 (OpenAI) and Claude (Anthropic) to generate step-by-step instructions for making a PB&J sandwich. The initial prompt was identical for both language models, instructing them to "generate instructions for making a Peanut Butter and Jelly sandwich". Following the LLM-generated instructions, we observed that Claude's output lacked essential details such as the absence of a plate, tablespoon, and instructions for opening the peanut butter and jam jars. Contradictions in steps and uncertainty about where to place the peanut butter or jam on the bread were also noted. Furthermore, there were no instructions related to health and hygiene. In contrast, ChatGPT produced complete instructions but overlooked health and hygiene best practices. Both models assumed prior cooking experience. Then one team member followed these instructions and made a PB&J sandwich with mixed results (see Figure 2).

Figure 2 One team member is following the instructions generated by the LLM.

To address these shortcomings, the team repeated the exercise with a more detailed prompt that "explicitly stated the absence of prior cooking experience and emphasized health and hygiene best standard practices". The generated instructions improved significantly under these conditions. The team then tried out the instructions of the different LLMs to see if they led to a nice sandwich.

5.2 Team 2: Visual instructions for making PB&J sandwich

Team 2 initially used DALL-E 2 to generate step-by-step visuals for making a PB&J sandwich, but the initial results were unsatisfactory, featuring repetitive images and nonsensical text. Attempts to refine the prompt led to equally unhelpful outputs, including a warning for potentially violating content policy and more images with irrelevant text. We then sought a different strategy, turning to ChatGPT-4 for a detailed prompt creation. This 181-word detailed prompt resulted in significantly improved images, though they still contained some peculiarities. Our team used these improved instructions to create a sandwich, encountering minor issues like attempting to combine the jars of peanut butter and jelly and placing jelly on the bread while it was still in the jar.

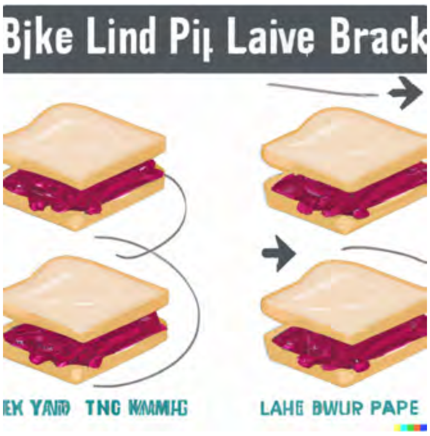


Figure 3 Attempt no 1: Would like step-by-step visual instructions for making PB&J sandwich.

Figure 4 Attempt no 1.5.

Then there were several more attempts to create great prompts for the individual steps. But it seemed to be too much work to do all the steps one picture at a time and thought to also combine instructions.

```
"Drawing of two pieces of bread on a plate"

A drawing of two pieces of bread on a plate, followed by a
    drawing of one piece of bread with jelly on it, next to a
     jar of jelly.

Would like step-by-step visual instructions for making a PB\&
    J sandwich without any text
```

Then the team went to CHAT GPT-4 to first create a prompt for DALL-E 2.

```
CHAT GPT-4: Generate a prompt for DALL-E 2 that will generate
     a set of images to illustrate how to make a peanut
    butter and jelly sandwich.
Create a series of detailed, colorful, and realistic images
    that visually guide through the process of making a
    classic peanut butter and jelly sandwich. The first image
```

Figure 5 Visual Instructions.

```
 should show a kitchen countertop with all the necessary
ingredients: a jar of creamy peanut butter, a jar of
grape jelly, a loaf of white bread, and a knife, all
neatly arranged with a clean background. The second image
 should depict the first step: spreading peanut butter
evenly on one slice of bread using a knife. The third
image should illustrate spreading jelly on the other
slice of bread, with the jelly glistening slightly to
show its texture. The fourth image should show both
slices of bread being carefully put together, with a
slight squeeze to make the peanut butter and jelly
combine slightly. The final image should present the
completed sandwich on a plate, cut diagonally, with a
small bite taken out to reveal the layers of peanut
butter and jelly inside, set against a bright, inviting
background. All images should be in high resolution and
showcase each step clearly for educational purposes.
```

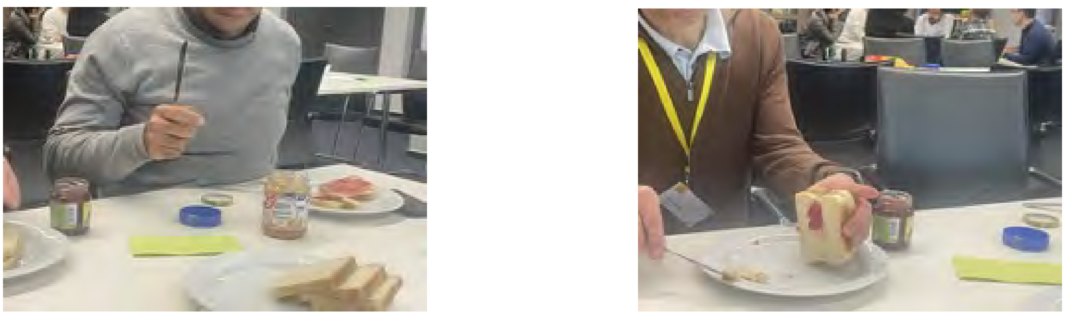


Figure 6 Two team members following the visual instructions.

5.3 Team 3: Textual Instructions Robot

The challenge in writing textual instructions was to decide on the assumptions as to the intelligence of the robotic arms + camera system. Do they understand instructions about color? Do they know what a piece of bread is?

Writing for a system with little intelligence proved to be too exhausting, so we (arbitrarily) assumed a mid-level of intelligence, e.g. – knowing what bread was, but not knowing how to open a jar.

The team wrote the instructions for 2 robots working together. Some fun moments occurred, such as when Robot 1 endlessly turned the jar lid, making the whole jar turn. We iteratively corrected the instructions to reach this outcome:

The 2 apparent errors resulted from turning the top slice of bread 180 degrees on the horizontal plane and not the vertical plane; and from placing the knife straight down.

One interesting dynamic observed was that robot 2, which had the job of holding the jars and slices so that they don't move while robot 1 did all the active stuff, seemed a bit depressed.

The instructions:

```
Robot 1 and 2:
Observe the items in front of you.

Robot 1:
hold BOTTOM half of jam jar with red and green lid

Robot 2:
Open jam jar with red and green lid, by twisting counter
    clockwise,
After 3 twists, place the lid on the table
Robot 1:
Hold slice of bread
Robot 2
pick up the other knife
insert the blade vertically half way into the open jam jar
Tilt the knife 90 degrees and lift the knife out of the jar
    vertically
Move the knife over one end of the slice of bread held by
    robot 2
Tilt the knife by 180 degrees so jam is facing down
Lower the knife until it touches the bread
Smooth the jam horizontally until it reaches the other end of
     the bread
Repeat steps 3 to 8 until the whole surface is covered in a
    layer of jam
Place down knife on table

Robot 1:
Let go of one slice of bread
hold BOTTOM half of  peanut butter jar with blue lid

Robot 2:
Open peanut butter jar with blue lid, by twisting counter
    clockwise,
After 3 twists, place the lid on the table
```

```
pick up one knife
insert the blade vertically half way into the open peanut
    butter jar
Tilt the knife 80 degrees and lift the knife out of the jar
    vertically
Move the knife over one end of the other slice of bread held
    by robot 2
Tilt the knife by 150 degrees so peanut butter is facing down
Lower the knife until it touches the bread
Smooth the peanut butter horizontally until it reaches the
    other end of the bread
Repeat steps 17 to 22 until the whole surface is covered in a
     layer of peanut butter
Put down knife

Robot 1:
Pick up a peanut butter slice of bread
Move it vertically so it is centered over the other slice of
    bread
Rotate it 180 degrees
Lower it onto the jam slice of bread so the edges of the two
    slices match
```

Figure 7 Well, somewhere there seems to be a little open question about removing the tools.

6 Ideation Session

In this ideation session we explored how Human-AI collaboration could unfold in various work domains. To envision and investigate various forms of human-AI interaction, participants used diverse prototyping techniques from low-fidelity proptotyping, to video prototyping, to generating sketches and images with AI.

6.1 Emily the Electrician

Abstract – zoom out from electrician to blue-collar work (in general)
Contributors: Tim Stratmann, Aaron Quigley, Sven Mayr, Susanne Boll, Larbi Abdenebaoui

Title: "Chronicles of AI Brilliance: A Design Fiction for Future Electrical Work"

With the growing influence of artificial intelligence (AI) the daily tasks of blue-collar workers will change. However, there is a lack of comprehensive frameworks or well-developed scenarios that effectively envision the seamless integration of AI within work environments and the subsequent interactions with human workers. We envision an integrated AI toolbox (both physical and digital) to provide support for daily tasks performed by blue-collar workers, with a key emphasis on bolstering overall performance (including safety, effectiveness, and efficiency), worker experience, and well-being at work. The vision outlined in this work proposes a collaborative landscape where AI acts as a supportive partner, amplifying worker capabilities while ensuring that humans retain control over the collaborative process except when safety and compliance to standards are at stake. Rather than displacing human agency, the objective is to create a symbiotic relationship, where the integration of AI enhances, rather than supplants, the role of blue-collar workers. This forward-looking perspective holds implications for the future trajectory of AI integration in blue-collar professions, highlighting the need for a balanced approach that prioritizes both efficiency and human-centric control as well as increasing well-being (by raising interest in the work).

Scenario Abstract: "A Day in the Life of Emily, an Electrician with an AI Toolbox (2058)"

Embark on a speculative journey to the year 2058, where the synergy between human intuition and artificial intelligence (AI) paints a vibrant tableau in the life of Emily, an electrician. The day unfolds with a ritualistic touch as Emily engages a vintage coffee machine, setting a nostalgic tone for the future. Emily's AI companion, aptly named EAI, orchestrates a symphony of seamless assistance. Picture an autonomous robot cart, adorned with a human-like robotic arm, gracefully retrieving tools selected by EAI. The morning spectacle continues with AR glasses projecting a nuanced dance of information, presenting a detailed electrical map juxtaposed with a simplified work overview at Emily's whim.

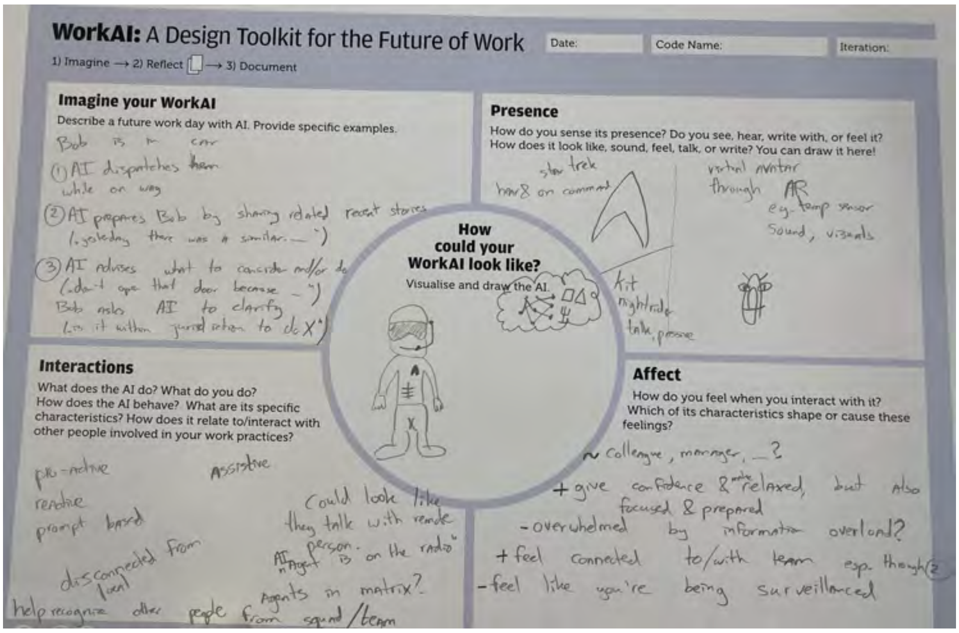
Figure 8 Emily refining her skills in the electrician in her AI-piloted electric car.

Figure 9 Emily the electrician in her automated vehicle preparing for the coming job.

As Emily commutes to a client's location, EAI initiates recurrent training, a ballet of skill refinement intertwined with the rhythmic hum of an AI-piloted electric car. The arrival at the client's site sees the EAI-piloted robot, a harmonious ballet of machinery, delivering tools to the workspace. A crescendo of innovation unfolds as Emily employs a scanning tool embedded in her phone, guided by the omnipresent EAI radar, ensuring both precision and safety. The narrative reaches its zenith as the EAI-piloted universal screwdriver choreographs an intricate drill, elevating the performance in wiring tasks. This design fiction unfolds as more than a mere speculative scenario; it encapsulates the harmonious collaboration between humanity and technology, offering a glimpse into the imaginative future where AI orchestrates a symphony of brilliance in everyday professional practices.

6.2 AI for Police

Contributors: Chris Janssen, Andrew L. Kun, Stefan Schneegass, Max L. Wilson



■ **Figure 10** Ideation about AI for police.

Artificial intelligence can help police to perform their job safely and effectively (see also Figure 10. Specifically, we envision that AI can help police in three ways:

1. **Use a variety of signals to predict events that require police response.** AI can use a host of signals to estimate the probability that there is an event where police are needed. This approach clearly presents pitfalls, as was shown effectively in the movie "Minority Report." The AI can incorrectly predict an event, and the subsequent police response can be harmful. Still, saving time in police response can save lives, so it is important to explore this possibility carefully.
2. **Use anecdotes to provide information.** While responding officers are traveling to the scene, the AI can prepare them by providing relevant information, including anecdotes from other officers who faced similar situations. Many first responders learn from each other through anecdotes. If AI can leverage this type of interaction with officers, it could fit very well into the way that they like to learn new information.
3. **Provide on-site support based on available data.** On the scene the AI can provide responding officers with context-relevant advice, such as where to look, what to look for, and how to act. The advice can be based on sensors that the officer wears, sensors in any police vehicles, and also on additional information, such as data about nearby mobile phones. Just as in the case of predicting events, this approach has many pitfalls. And just as in that scenario, there are significant possible benefits, which means that it is important to explore this possibility as well.

6.3 AI for Education / Teaching – “Social Synergy”

Contributors: Marta Checchinato, Duncan Brumby, Anna Cox

The presentation discussed a futuristic AI system designed for higher education that aims to personalise teaching materials and assessments to enhance and personalised learning experiences. This system includes a desktop interface that provides educators with information on student needs, such as disabilities and language preferences, to aid in teaching delivery and facilitate social interactions. The AI could also function like a dating app, pairing students for group work based on their preferences to foster personal growth and a sense of belonging. We also considered the use of tools like Microsoft’s Speaker Coach to reduce bias and discrimination while being mindful of potential concerns about surveillance and privacy. Similarly, Harvard-tool called “Teachly” provides interesting technology to consider that analyses classroom seating data to provide feedback on teacher-student interactions and potential biases. The AI would provide live, real-time feedback and input to the teacher.

6.4 AI for Theater – “Alladdin AI”

Contributors: Marios Constantinides, Michal Rinott, Shadan Sadeghian, Orit Shaer, Erin Solovey

Figures 11 to 14 show extracts from the video produced by the contributors during the seminar.

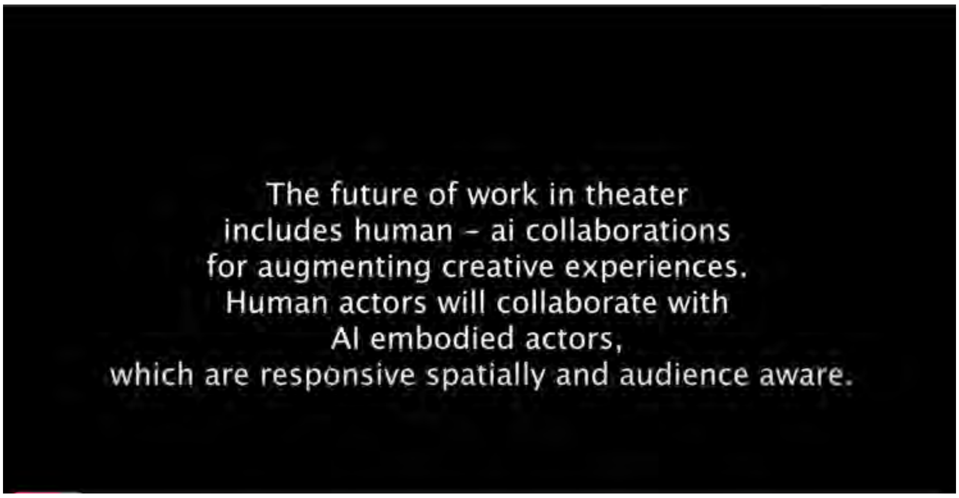


Figure 11 Envisioning human-AI collaboration for Theater.

Figure 12 Envisioning Human-AI collaboration for Theater: Production of Alladdin-AI.

Figure 13 Envisioning Human-AI collaboration for Theater: human actor scans the stage in preparation for human-AI collaboration.

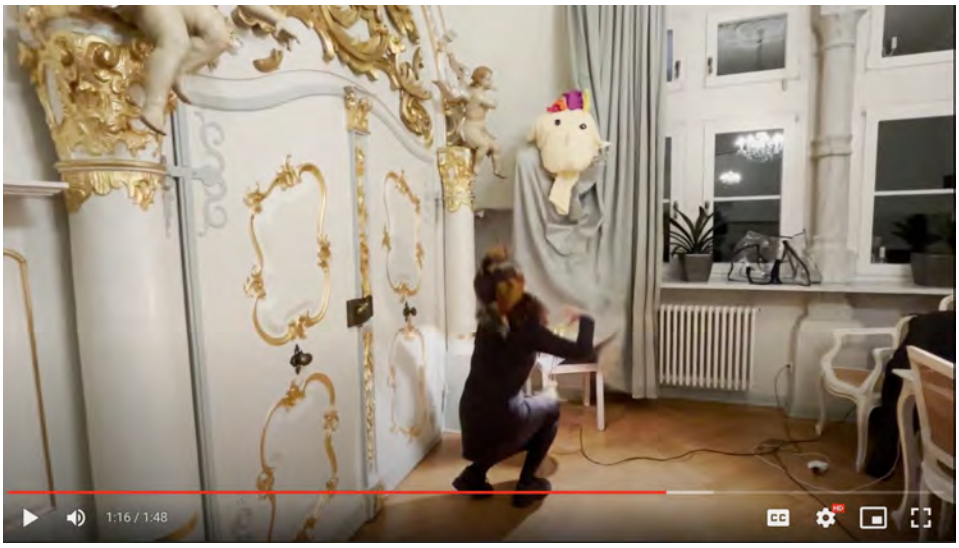


Figure 14 Envisioning Human-AI collaboration for Theater: human and embodied AI co-acting.

7 World Cafe: Topics in human-AI interaction for work

In this session, participants discussed four themes of human-AI interaction for work in the format of a world cafe. The themes were identified from brainstorming among the participants.

Participants first produced a large number of research themes and facets of human-AI collaboration and interaction for work and then voted on which themes to further explore through discussion in this session.

Each theme was assigned a moderator (table head) and then explored around a table located in different areas of the seminar room. Participants rotated between the different themes while the table head led the discussion and summarized the results. The following themes were discussed:

7.1 What are interaction models for efficient/good human-AI collaboration?

Table head: Philippe Palanque

This was the starting point of the discussion based on input from participants during the identification of topics relevant to the seminar.

At this round table we started by slicing the title and discussed each of the following aspects: What is a model? This question was addressed by making explicit the difference between "Model with big M" and "model with small m". Model is a set of concepts and relationships between concepts (e.g. the Entity/Relationship diagram [Chen 76]) while a model is a description of a part of the real world using concepts from a Model.

Is it reasonable to make a difference between AI and automation? It looks like AI is only a means to implement automation and issues related to automation seem fully relevant. It might be relevant to look at work in IFAC (automation and control) and SIGDA (Design Automation).

What are the qualities of a model? Following the discussions, we discussed the qualities of a Model and the qualities of a model. Efficient, and good were in the title. Other properties were identified such as reliability, and dependability, ... as for model properties such as expressiveness, coverage concerning the items to be modeled, coverage of the concepts in the Model, and the need for representing important and relevant real-life elements to be captured in the model, ... What do we mean by collaboration in the title? Discussions went to the allocation of work between AI and the human with different organizations in mind (H-AI, H-H-AI, H-AI-AI, and the difference with H-H-H collaboration).

Workload: a perspective for human and AI The discussion progressed toward the definition of "good" design and assessment of good designs including assessment of workload. For the human side the Yerkes Dodson law was mentioned and discussed and then what the bell curve would look like from the AI perspective

For the AI perspective overload would be (for instance) decreased availability of the service while underload would be low input to the continuous training of the model. Collaboration perspective between AI and Human Discussions evolved around the notion of human-human computer-mediated collaboration such as the Clover model [1]. The importance of slicing collaboration into Production (output of the work), Communication (between collaborating humans), and Cooperation (organizing and allocating work to entities) was useful in discussing collaboration with automation.

References

1 Yann Laurillau and Laurence Nigay. 2002. Clover architecture for groupware. In Proceedings of the 2002 ACM conference on Computer supported cooperative work (CSCW '02). Association for Computing Machinery, New York, NY, USA, 236–245.

7.2 How to design for meaningful work?

Table lead: Shadan Sadeghian

We all spend a large proportion of our lives working. In its original definition, work is a way of earning a living. However, even though people (ideally) have several ways to fulfill their financial needs, they choose one over the other, and sometimes even the one that pays them less! This raises the question of why do we work. One answer is that besides being a way to earn a living, work can be a source of meaning in life.

Of course, we are not the first to raise this question. Over the past decades, several researchers have looked into work meaningfulness and characteristics that make a job meaningful. An example that was raised was the model by Hackman and Oldham [1] which defines five job characteristics that derives motivation at work:

- Skill Variety: The degree to which a job requires various activities, requiring the worker to develop a variety of skills and talents.
- Task Identity: The degree to which the job requires the jobholders to identify and complete a workpiece with a visible outcome.
- Task Significance: The degree to which the job affects other people's lives. The influence can be either in the immediate organization or in the external environment.
- Autonomy: The degree to which the job provides the employee with significant freedom, independence, and discretion to plan out the work and determine the procedures in the job.
- Feedback: The degree to which the worker has knowledge of results. This is clear, specific, detailed, actionable information about the effectiveness of his or her job performance.

In our group, we delved into the potential impacts of future AI collaboration on various aspects and, consequently, on the perception of job meaningfulness. A central and recurring theme in our discussions was the recognition that while current workplace technologies primarily focus on enhancing productivity, as often gauged in Human-Computer Interaction (HCI) by measures of effectiveness and efficiency – such as error reduction and improved outcomes – these factors alone do not encompass all elements contributing to job satisfaction.

A fundamental source of job meaningfulness is the sense of competence derived from possessing a diverse array of skills relevant to one's job. Collaborating with AI may involve delegating certain tasks to the AI itself. If these tasks are integral to the individual's sense of meaningful work, how can we sustain this feeling of significance? Alternatively, what new skills should individuals cultivate to enhance their job satisfaction? Can the design of human-AI interaction play a role in augmenting feelings of competence among humans?

We also delved into the societal dimensions of work, particularly exploring the dynamics of relationships among colleagues. Numerous topics within this sphere were considered. For instance, questions arose about social interactions such as "Whom do I join for an after-work beer if my colleague is an AI?" or pondering whether success achieved while working with AI in a team would instill a sense of pride in the collaborative accomplishment. While the integration of AI in work practices might enhance performance in the short term, there's a concern that over the long run, it could erode feelings of connection and belonging, subsequently diminishing motivation in the workplace. This prompted us to question, "Can AI be designed to function as a good colleague, and if so, how?"

Another societal consideration involves workers' rights. Currently, the allocation of tasks between human workers and AI is primarily determined by system designers (developers) or organizational management. Unfortunately, workers often have minimal, if any, input into what tasks they perform or how they carry them out in their jobs, despite being the most affected group. This lack of involvement can potentially decrease their job satisfaction, as they might be assigned tasks that the AI cannot handle or be required to spend long hours monitoring AI activities. This raises the question of how we can design interactions with AI in the workplace to preserve or even enhance feelings of autonomy and uphold workers' rights in decision-making processes.

References

1 Hackman, J. Richard, and Greg R. Oldham. "Development of the job diagnostic survey." Journal of Applied psychology 60.2 (1975): 159.

7.3 How to design for equitable human AI collaboration?

Table lead: Neha Kumar

We began by discussing the topics that were brought up by participants around equity in human-AI collaborations. We asked first, "Who are we equating? Is it humans to other humans, or humans to AI?" Realizing that in different contexts, the challenges arising might be different, we decided to look across the Design, Build, Use, and Iterate stages to ask what challenges these might entail, and what the hard questions might be. For example, if we were to consider an LLM, what key questions should we be asking in the design, building, use, and iteration of an LLM. In other words, what might the "CV" of an LLM look like, and what questions must it contain answers to? E.g. where do its origins lie, how did it get to where it is now, what are its strengths and limitations, and how does it learn and grow? To understand the "learning and growth" we would need to lay out many key priorities as well.

7.4 What are the skills humans and AI need for successful collaboration?

Table lead: Christian P. Janssen

At this round table, we addressed the following main question: "What are the skills humans and AI need to have for successful collaboration?" We started with a discussion of "How does one even capture or define skills?" Do we need different terms or definitions for human-AI collaboration compared to human-human collaboration? For example, is a skill like "interpersonal skills" – which is maybe mostly used for human-human collaboration, similarly defined for human-AI collaboration? See Table 1 about what is known about skills for collaborations between two types of partners:

Table 1 Skill collaborations between two types of partners.

		Partner 2	
		Human	AI
Partner 1	Human	(within reason) well described	
	AI	To be studied	Defined by former protocols

We then talked about what skills are required for humans, for AI, or for the interaction (Table 2):

Table 2 Required skills for humans, AI or for the interaction.

Human	**AI**	**Interaction**
Knows or defines the "value" or does "curation"	Decomposes or classifies situation to (better) communicate the problem	Make achieving the goal easier
"Steers" conversation	Creates proposals	Task "translation" and "evaluation"
Evaluates	Expresses limitations or uncertainties on their recommendations	Norms / expectations / appropriate actions grounded
		Added value of collaboration is made explicit

What's needed for human-AI collaboration? Two additional aspects were identified:

1. Probably, there's some form of alignment needed.For human-human communication alignment is sometimes implicit: You nod your head, you look at a shared artifact. By contrast, in human AI interaction, alignment needs to be made more explicit. The understanding among the groups was that alignment in human-AI collaboration is currently done well.
2. Common sense and satisficing (instead of optimizing). Sometimes an AI should slightly "break the rules" to achieve a result. For example, if you are in a rush and it is relatively safe to do, drive slightly above the speed limit to make it to an urgent appointment.

With the first two groups, we talked quite a bit about the role of uncertainty. Some uncertainty or some probabilistic dynamics are probably useful in an AI system, because it creates, for example, creative discussions with large language models. Due to some variations, it's not always the same thing. At the same time, having too much uncertainty or probabilistic variation can be bad.

Given that there is some level of uncertainty, there is a question of "how do you even communicate uncertainty to the user?" and "how does this differ between different users?" For example, how does a child or novice handle uncertainty versus an adult or expert user (that perhaps knows better how to contextualize the uncertainty)?

We also discussed that for human-AI collaboration, the required skills can change at two levels:

1. The job / task / goal / life itself can change due to AI / Automation. These are in some sense the "bigger questions" that are about potentially radical changes to work.
2. For interaction with the AI. These are mostly about specific tasks/interactions "how can I achieve X with the new system".

We also discussed AI low literacy. As not everyone might be (immediately) able to work with novel forms of AI. Therefore, systems require some robustness, and allow some exploration capability for the user with the option to "undo".

Finally, we discussed organizational contexts. In some organizations, workers might be forced to use a specific system. A structure is then needed for (re-)training. But how? Is training offered generic to a group? Or customized to an individual?

How does one get over local maxima? That is, if one found a useful routine for approaching a task, but new / different tools have the potential to improve your performance – how do you get people to try out these techniques in a safe way?

This probably requires some organizational change. Time to "catch up" and "invest" in new technology is needed and should be appreciated by organizations and society.

8 Working groups

8.1 Forms of interaction

Contributors: Orit Shaer, Mohit Jain, Andrew L. Kun, Mihal Rinott, and Bastian Pfleging

In this session, we discussed different aspects of envisioning how people will use and interact with AI technology in the future. Could we cross between interaction paradigms and current+future AI uses to create a matrix that will show existing and new potentials for interactions with AI?

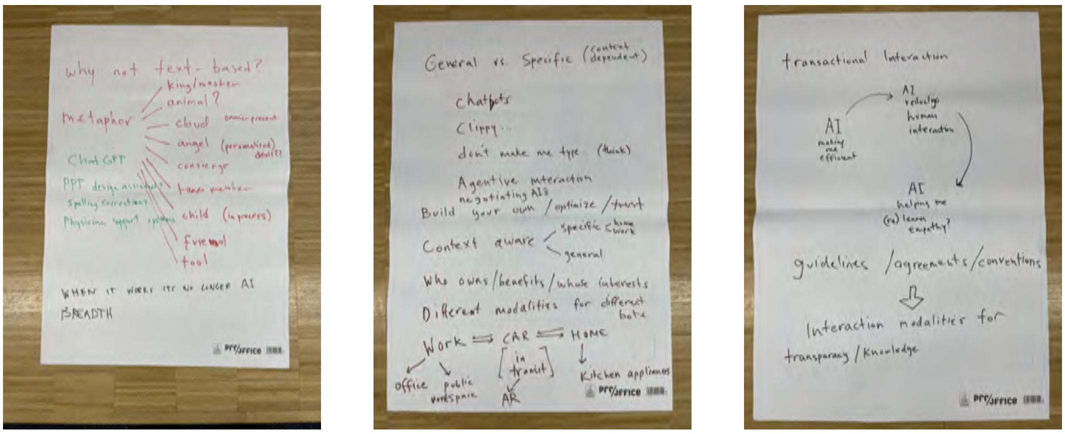


Figure 15 Notes for forms of interaction discussion.

Divergent vs. convergent AI Systems. We discussed the divergent (many AIs) versus convergent (one AI to rule them all) possible futures of interaction with AI. Convergent AI would mean the persistence of a godlike AI entity across contexts, such as Home → Car → Work, requiring a move between different interaction paradigms. Multiple AIs represent a more diverse view of technology. Ownership and interests could be more clear and transparent. AI Systems Supporting Users: How will different user systems interact with each other?

8.2 AI for science

Contributors: Anna Cox, Christian P. Janssen, Aaron Quigley, Max L. Wilson, Stefan Schneegass, Susanne Boll, Phillippe Palanque, Tim Stratmann, Sven Mayer, Marios Constantinides
Summarized by: Anna Cox, Christian P. Janssen, and Max L. Wilson

We first talked about how the bigger scientific process is structured. One proposal is that you can talk about "science missions", which are an interactive process of ideation, discovery, and commercialization/impact. Then, who are we talking about when talking about scientists? Are these academics? Or Industry researchers? Or creatives/knowledge workers? If we are talking about academics, are we talking about their scientific activities or their everyday work (e.g., balancing research, teaching, and service)? We considered the following four factors affecting the future of science with AI:

AI for the Scientific Process versus AI for the Work of Scientists. We consider that there will be two impacts. Many tools are being developed to help achieve specific tasks in the scientific process, but AI will affect other forms of work that scientists do (time management, meetings, reviewing), and indeed are done in the infrastructure that creates science (finance, administration, etc).

- AI advances will enable people to do more science, and to do science in new ways.

AI Speed-ups versus Experiencing the Process. We consider that AI will automate and complete many tasks for people that will enable science to happen much faster – but what will we miss? Is it important, for learning and understanding, to be "in the loop" of the tasks AI will complete? Importantly for the rigor behind science, we need to see into the process and know that it can be trusted. This is true even if AI is holding us as researchers to take rigorous processes. This includes regulation about how and where AI is used.

- AI must enable science, but we will need to retain scientific rigor.

New Skills and Changing Skills. We consider that AI will have the so-called "Google Maps" effect on what skills we have, and what skills people will develop in life. Does it matter that we have lower attention spans due to social media? Do we have new skills of consuming knowledge in faster and better media formats? We wonder what important skills in life will be lost as we become comfortable with AI doing things we do not need to do. We wonder further when the impact of these losses will be discovered (when will we realize that attention span had a critical impact).

- We need to understand what AI literacy is, and what the future landscape of human skills will be.

One AI (eg a digital assistant) versus Specialized Tools in a Toolbox. We consider whether the AIs around science will be integrated into enabling all the science we are doing, or is being done within one organization. Or if these are each specialized tools that we use to achieve our goals in the science process.

- We will need to study the introduction of AI into the science process, for how these evolve and whether they become integrated.

8.3 Roles and metaphors

Contributors: Shadan Sadeghian, Marta Cecchinato, Naveena Karusala, Neha Kumar, Naomi Yamashita

Throughout history, humans have employed various strategies to navigate the unfamiliar and cope with uncertainties. We've assigned names like aliens or zombies to unknown entities, categorizing them, or utilized metaphors to apply familiar interaction strategies. Even in our interactions with technology, such strategies prevail. We've adopted metaphors such as desktops and files to understand and engage with technology. As AI becomes more prevalent in our daily lives, similar tactics are employed. AI systems, with their unique capabilities like agency, opacity, and anthropomorphism, are introduced as new entities. We often seek familiar characters in these unfamiliar beings, describing our relationships with them as master-apprentice or co-pilot. However, these metaphors not only define the relationship but also shape the interaction dynamics and notions of accountability when things go awry. Hence, it's crucial to carefully consider the metaphors we use to make the unfamiliar familiar and determine when a specific metaphor is beneficial for describing AI interactions.

Drawing on Social Practice Theory [1], our discussion delved into the skills, material, and meaning sought through technology interactions in various work practices. We concluded that these elements heavily influence the choice of metaphors used to define interactions. For instance, the metaphor of a "mask" for a real-time translator AI can both mask language skill flaws positively by presenting fluency and negatively by filtering out unique expressions, thus diminishing diversity. We ended the session by posing research questions such as: Why should we rely on metaphors, and where do we draw the line between familiarizing entities and neglecting their unique characteristics? Which metaphors are suitable for integration into specific practices? How do technology attributes and social settings shape these metaphors and subsequently influence the roles and relationships between humans and AI?

References

1 Yann Laurillau and Laurence Nigay. 2002. Clover architecture for groupware. In Proceedings of the 2002 ACM conference on Computer supported cooperative work (CSCW '02). Association for Computing Machinery, New York, NY, USA, 236–245.

9 Prediction of Future Scenarios

One of the seminar goals was to discuss important challenges/research problems, and accompanying hypotheses. To envision how human-AI interaction for work could evolve in the coming years, in this session we asked our participants to discuss (in groups) potential future scenarios for the next decade and to think of very utopian, but also very dystopian versions of our future of work and AI.

We wanted them to find provocative ideas, which do not necessarily need to be right. Each group should come up with at least 3 utopian scenarios and at least 3 dystopian scenarios. Furthermore, we asked them to revisit their utopian scenarios, to understand what could go wrong, i.e., lead towards the dystopian scenarios? In addtion, they should think about which user groups will be affected and how. Similarly, we asked to also revisit the dystopian scenarios to investigate what could be done to fix the situations and understand which user groups will be affected and how.

Based on the domains addressed in previous sessions, we subdivided into four groups that developed scenarios for the following contexts: *Healthcare*, *higher education*, *workers' well-being*, and *manufacturing*. The following subsections outline the results from these group discussions.

9.1 Utopian and Dystopian views of work, health, and AI

Contributors: Max L. Wilson, Aaron Quigley, Marios Constantinides, Marta Cecchinato, Naomi Yamashita

The working group considered 6 scenarios, each could have been written as utopian or dystopian, and in the end their choices could have been reversed easily.

9.1.1 Healthcare: Utopian scenarios

Figure 16 provides a summarizing overview of the utopian scenarios and what could go wrong. The scenarios were the following:

- **AI replaces General Practitioner (Personal/Family Doctor)** In the next decade, we predict that... People will use AI to perform the function of the initial medical triage currently undertaken by a family/personal doctor (or General Practitioner)
- **Autonomous ambulances (or Ambulance Drivers)** In the next decade, we predict that... The wait times for ambulances will be eliminated by self-driving cars
- **AI provides personal proactive dynamic healthcare delivery** In the next decade, we predict that... New AI-enhanced sensing devices will proactively support individual health. Healthcare will go through a new transformation from being preventive (fixing problems as they come) to being proactive (anticipating problems).

9.1.2 Healthcare: Dystopian scenarios

Figure 17 provides a summarizing overview of the dystopian scenarios and how these issues could be addressed. The following scenarios were discussed:

- **Empathetic Companion for Carehomes.** In the next decade, we predict that... New empathetic robots will be introduced to care homes for the elderly, which decrease the human contact that they experience.
- **The Hilton Hospital.** In the next decade, we predict that... the Hilton hotel chain will announce a lifetime platinum membership for anyone who signs up for their "Healthy Hilton program".
- **Robot nurses.** In the next decade, we predict that... nurse staffing levels will reach catastrophically low levels as overworked nurses quit the field en masse, and aspects of their work will be replaced by care robots.

9.2 Utopian and Dystopian view on Higher Education

Contributors: Christian P. Janssen, Erin T. Solovey, Stefan Schneegass, Sven Mayer, Naveena Karusala

In the next decade, we predict that...

- ... Machines play a central role in teaching. As a student, you use a machine to get things done. You do not need to learn to calculate, write, think (?), code.
  - Context:
    * Different students work on different things. Lecturers and school context need to adapt to aid where needed.
  - Upside / Utopian:
    * personalized learning (intelligent tutoring);
    * no exams might be needed as system keeps track of what you know and don't know. Formative tests.

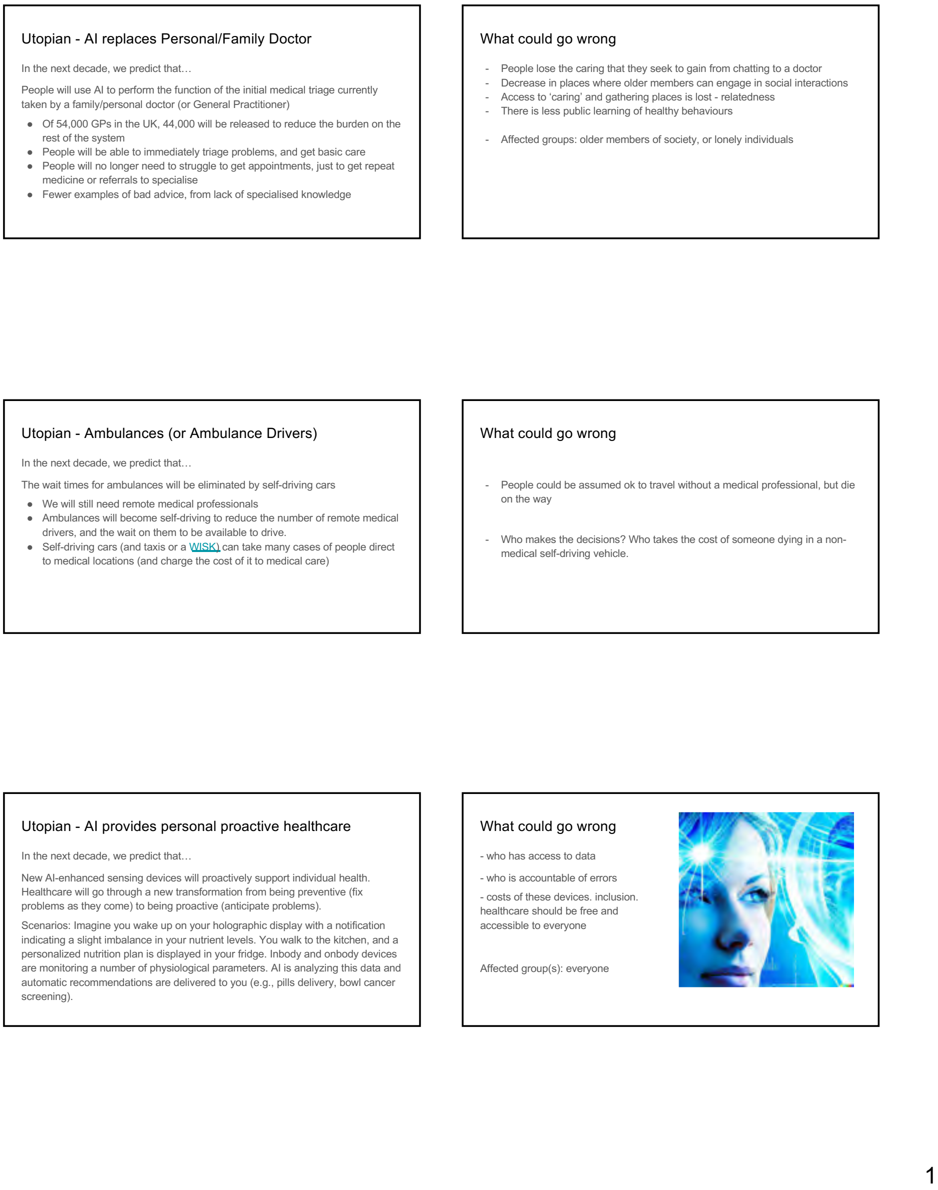


Figure 16 Presentation slides: Utopian scenarios – Healthcare.

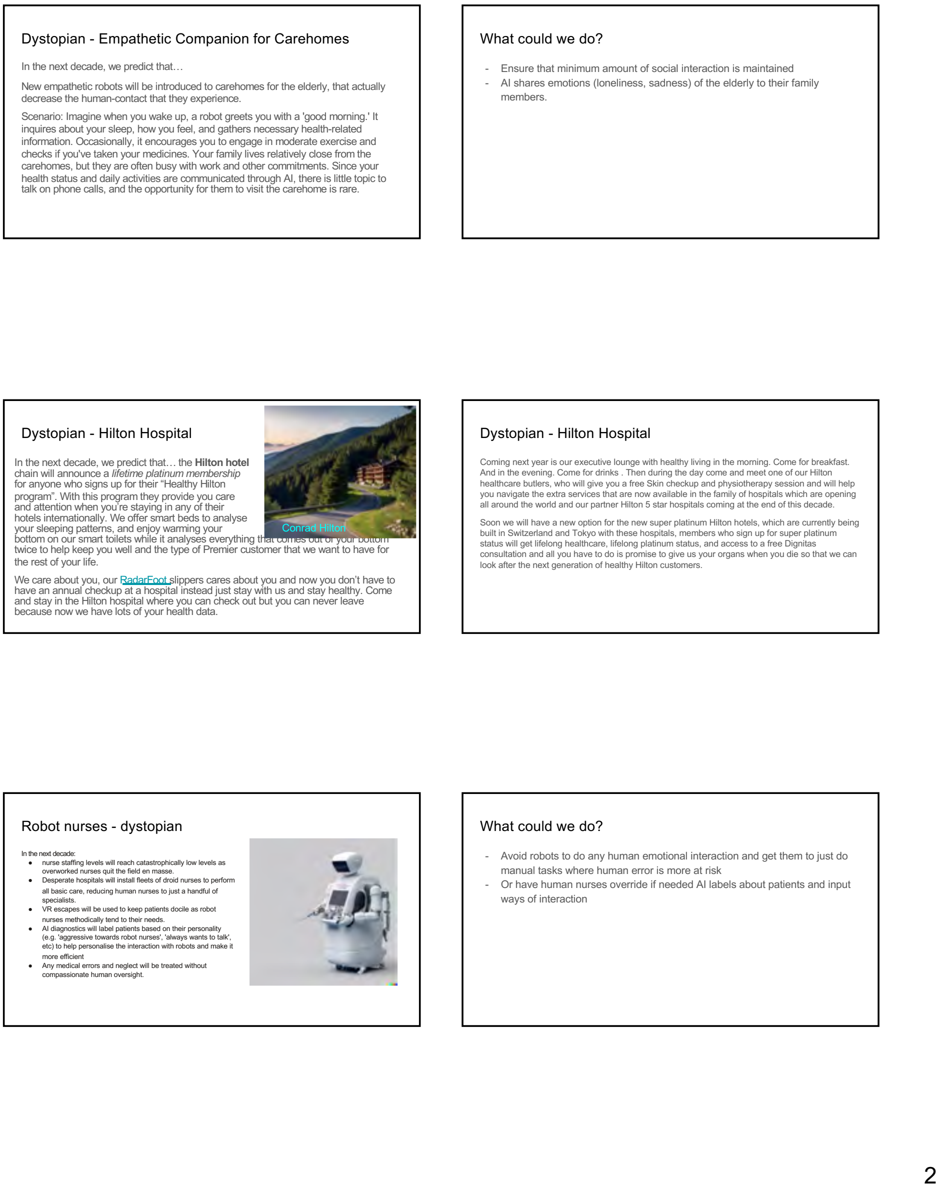


Figure 17 Presentation slides: Dystopian scenarios – Healthcare.

    - school might not be "boring" as calibrated to an individual's interest and learning style
    - university can "guarantee" a specific quality level
    - the teacher has the time to (also) focus on the individual student (with the help of AI).
    - AI can see what types of errors students make beyond typical cases. AI can use those responses to suggest different examples, material, and instructions for rare cases.
    - classrooms are multi-modal: AR and VR usage
  - Downside / Dystopian
    - when code does "wrong" things, knowledge is not there on how to "correct"
    - No room to "bend the rules" as students, and try different things. "Eenheidsworst" / "einheitsbrei"
    - LLMs / Machines create the lectures and assignments, tenured staff is not needed – their creative input and context is not needed.
    - Teacher needs to focus on many things as well due to range of skills / levels / interests of students.
    - Human-human interaction might be missing (or reduced)
    - "Sorting hat" (from Harry Potter) for teaching sorts pupils early into categories of where they should flourish. "You should do STEM / Humanities / ...", and their path is set for them. No broader education anymore
  - Mitigating factors:
    - Ensure that human-human teamwork is part of the curriculum
    - Ensure that students are challenged on topics outside of their "bubble"
    - Include assignments that train "basic" skills without the AI / LLM /... such that students still know
- Universities no longer exist ?!?!?! (or... many close / change / restructure)
  - Context: in general because...
    - "teaching is done by the AI"
    - "research is done in industry"
    - PhD training is done with the AI or while working in industry
    - Specific types such as "community colleges" or "research universities" might change tremendously. E.g., community colleges have current research universities to teach; Research universities are merged with industry.
  - Upside / Utopian
    - Lots of NGOs, and research institutes do practical research that has societal value.
    - Concepts of ethics and social good are shared more easily and widely
    - No crapshoot "research for money" for industry needs to be done by academics
    - No "learning factories"
    - CS is not disappearing??
  - Downside / Dystopian
    - The "Sillicon valley university" is blend of current universities and industry
    - (research) (teaching) "Agenda" is set by external parties
    - Secret stealing of ideas; or inserting of bad AI (because everything is digital / online)
    - Global competition / war as international norms do not align with local culture and people do not meet to "align"

    - “Exploration trip in 2050” will visit the ruins of Dagstuhl to see how people communicated at the start of the 21st century. (like how we went to Völklinger Hütte)
    - Only crapshoot “research for money” for industry needs to be done by academics
    - Some fields of academia disappear (e.g., “no money for humanities”)
  - Mitigating factors: Ensure a place for fundamental/risky research within universities (as industry might not want to try)
- AI allows us to do research that used to be “impossible” like solving DNA sequences, or cure cancer…
  - Context:
    - AI allows advanced pattern recognition beyond human capabilities
    - AI allows the integration of information
  - Upside / Utopian
    - Individualized healthcare; doctors who actually can do research (instead of case studies)
    - Democratization of research: doctors and practitioners have access to a team of (AI?) researchers to help with solving difficult questions
    - Faster research output
    - (reasonably good) Digital Twin or testing platform of whatever system/context you study. Model-driven research. Less need for animal studies, human studies, …
    - Integration of theory and insights. “Unified theory of X”
  - Downside / Dystopian
    - Costly research: run a server for “trivial” research question
    - Quality control: if AI does the research, how will you know what the output is/means and how to validate it?
    - AI hallucinations: making stuff up that sounds plausible to human
    - More pressure to do research fast; less emphasis on quality. More work created because no time needed for research.
    - Less room for divergence; when it is there, it might give grounds for “fake news” and alternative realities/truths
    - 1984?
  - Mitigating factors:
    - Tools to check AI
    - Competition/variety of AIs / Algorithms
    - Open science
    - Certification / Peer-review like systems
    - Human-in-the-loop (to make it slow?)
    - Alt.chi like venues that give a place for alternative perspectives but are at the same time not rabbit holes for radicalized ideas.
- How do you evaluate faculty in age of AI?

9.3 Wild Predictions on Workers' Wellbeing

Contributors: Tim C. Stratmann, Larbi Abdenebaoui, Mohit Jain

9.3.1 Utopian scenarios: Workers' Wellbeing

In the next decade, we predict that...

1. workers have more freedom in choosing their task
   - more flexibility
   - more meaningful work
2. there is no hard physical work
3. AI can help to have an ideal working relationship among employees through perfect communication

   How can these utopian scenarios turn into dystopian ones?

1. No relatability, Humans cannot be trusted, AI learns to untrust Humans → only unimportant jobs, meaningless work, Black Mirror "Cycling for Energy"
2. All Physical Work is done by Robots → fat, lazy, loose cognitive abilities, → become a "Surrogate"
3. Manipulation and superficial relationship, no free speech, no unions any more → avoiding conflicts but, conflicts are also important, robbing society from meaningful conversations, company → political manipulation

9.3.2 Dystopian scenarios: Workers' Wellbeing

In the next decade, we predict that...

1. humans will work for the AI, Matrix
2. humans will lose every working skill. → we cannot repair, innovate
3. a permanent presence of AI – "Angel/Daemon" advising workers all the time

What can be done to fix these dystopian scenarios?

1. AI is democratic and willing to do the best for human sustainability → Perfect World. Optimal use of resources, care for the workers.
2. Gain the skill to interact with AI systems → learn new things on demand and rapidly, Matrix
3. Control the level of monitoring, including the possibility to turn it off → get the support we want when we want.

9.4 Utopian and Dystopian Predictions for Manufacturing

Contributors: Michal Rinott, Shadan Sadeghian, Anna Cox, Duncan Brumby, Philippe Palanque, Neha Kumar

This group discussed scenarios related to manufacturing. Before diving into different scenarios, the group raised the question whether utopia and dystopia are opposites and whether these opposites connect? In their discussion, it turned out that there could be a continuous cycle between such scenarios, which brought the connection to the word Saṃsāra: Saṃsāra is a Pali/Sanskrit word that means "wandering" as well as "world" wherein the term connotes "endless cyclic change".

This lead to the following Samasaric scenario with different steps and situations:

- As manufacturing increased its automation, many factories reached the "lights off" stage where the last worker was fired.

- People went home to unemployment. The world continued to pile up with manufactured goods.
- At home, people became bored, anxious and depressed.
- They wanted meaning. They wanted jobs back, but not exactly like before. This lead to something like "Slow Manufacturing"
- Amazon, Alibaba, and others made a 180-degree turn and decided to save the earth rather than get rich. They were rich enough now and hoping to save the earth for their children. Imagine: "Customers who bought this item did not buy anything else for a week."
- Big manufacturing factories, experiencing a consumption crisis, changed: moving to locations with renewable energy, such as warm sunny locations, to cut energy costs
- Manufacturing for personal goods increasingly moved to a distributed model of local manufacturing: at or near home.

  AI was used to optimize this process, making the manufacturing of some goods local and community-based, with each participant manufacturing a part of the whole product.
- Following the model of Crowd Cow, even farming was now subsistence-based: done in home gardens with AI helping to plan and optimize yield at the community level.

Alas, the cycle of samsara may continue, with some people wanting more... and thus communal manufacturing may once again be replaced by large corporations...

And so the cycle may continue between utopia and dystopia.

10 Summary and Next Steps

Participants of Dagstuhl Seminar 23452 "Human-AI Interaction for Work" spent a productive week discussing a variety of topics related to human-AI interaction for work. We focused on understanding where this broad field stands now. We also worked toward understanding both the positive and negative outcomes that are likely to be the results of incorporating AI into work. This included sharing our own work with each other, as well as discussing our own understanding of the relevant scientific literature (which is indeed very broad). It also included turning our attention to art (movies, TV shows, and books), as well as popular science books, to gain additional insight into problems and possible solutions.

We wrapped up the seminar with a session discussing our individual and collaborative next steps. Many participants are contributors to the CHIWORK symposium series – seminar participants serve on the CHIWORK steering committee, they are organizing the 2024 event in the UK, and they are contributing papers and reviews. Multiple participants are also planning to include AI-for-work topics in the IEEE Pervasive magazine, both through an upcoming special issue, and through a regular column in the magazine. We also discussed multiple collaborative research ideas, many of which focused on using LLMs for work. Finally, in all of our plans, we were determined to heed the warning of the abandoned Völklinger Hütte steel plant, which was our destination for the customary Wednesday afternoon field trip – we want to build tools that will allow people to flourish, and avoid building tools that dehumanize work.

As organizers, we would like to wrap this document up by expressing our gratitude to all those people who contributed to the success of this seminar. First and foremost, we would like to thank the team at Schloss Dagstuhl. Their dedication and their warm hospitality were evident in all of their actions, from organizing the meeting, to hosting us at the castle. And of course, we are most grateful to the seminar participants who took an entire week out of their busy schedules to join us in order to create new scientific knowledge in the field of human-AI interaction for work.

11 Annotated Bibliography

References

11.1 Understanding Work and productivity

1 Sandy J. J. Gould, Anna Rudnicka, Dave Cook, Marta E. Cecchinato, Joseph W. Newbold, Anna L. Cox, Remote Work, Work Measurement and the State of Work Research in Human-Centred Computing, Interacting with Computers, Volume 35, Issue 5, September 2023, Pages 725–734, https://doi.org/10.1093/iwc/iwad014

2 Joseph W. Newbold, Anna Rudnicka, David Cook, Marta Cecchinato, Sandy Gould and Anna L Cox (2021): The new normals of work: a framework for understanding responses to disruptions created by new futures of work, Human–Computer Interaction, https://doi.org/10.1080/07370024.2021.1982391.

3 Yang, L., Holtz, D., Jaffe, S. et al. The effects of remote work on collaboration among information workers. Nat Hum Behav 6, 43–54 (2022). https://doi.org/10.1038/s41562-021-01196-4

4 Christian P. Janssen, Stella F. Donker, Duncan P. Brumby, Andrew L. Kun, History and future of human-automation interaction, International Journal of Human-Computer Studies, Volume 131, 2019, Pages 99-107, ISSN 1071-5819, https://doi.org/10.1016/j.ijhcs.2019.05.006.

5 Brumby, D.P., Janssen, C.P., Mark, G. (2019). How Do Interruptions Affect Productivity?. In: Sadowski, C., Zimmermann, T. (eds) Rethinking Productivity in Software Engineering. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4221-69

6 Nyholm, S., Smids, J. Can a Robot Be a Good Colleague?. Sci Eng Ethics 26, 2169–2188 (2020). https://doi.org/10.1007/s11948-019-00172-6

7 Mark, G., Kun, A. L., Rintel, S., and Sellen, A. (2022). Introduction to this special issue: the future of remote work: responses to the pandemic. Human–Computer Interaction, 37(5), 397-403.

8 Illing, J., Klinke, P., Pfingsthorn, M., and Heuten, W. (2021). Less is more! Support of Parallel and Time-critical Assembly Tasks with Augmented Reality. In Proceedings of Mensch und Computer 2021 (MuC '21). Association for Computing Machinery, New York, NY, USA, 215–226.

9 Kun, A. L., Shaer, O., and Iqbal, S. (2021). The future of work: COVID-19 and Beyond. IEEE Pervasive Computing, 20(4), 7-8.

10 Kun, A. L., Sadun, R., Shaer, O., and Teodorovitz, T. (2020). Where did the commute time go?. Harvard business review.

11 Orii, L., Tosca, D., Kun, A. L., and Shaer, O. (2021, September). Perceptions of trucking automation: Insights from the r/truckers community. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 137-146).

11.2 Guidelines and Frameworks

12 Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. https://doi.org/10.1145/3290605.3300233

13 Philippe Palanque. 2020. Ten Objectives and Ten Rules for Designing Automations in Interaction Techniques, User Interfaces and Interactive Systems. In Proceedings of the International Conference on Advanced Visual Interfaces (AVI ’20). Association for Computing Machinery, New York, NY, USA, Article 2, 1–10. https://doi.org/10.1145/3399715.3400872

11.3 Explainable AI

14 Sven Schultze, Ani Withöft, Larbi Abdenebaoui, and Susanne Boll. 2023. Explaining Image Aesthetics Assessment: An Interactive Approach. In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval (ICMR ’23). Association for Computing Machinery, New York, NY, USA, 20–28. https://doi.org/10.1145/3591106.3592217

11.4 AI in the workplace

15 M. Constantinides and D. Quercia, “Good Intentions, Bad Inventions: How Employees Judge Pervasive Technologies in the Workplace,” in IEEE Pervasive Computing, vol. 22, no. 1, pp. 69-76, 1 Jan.-March 2023, doi: 10.1109/MPRV.2022.3217408. keywords: Artificial intelligence;Ethics;Productivity;Employment;Task analysis;Psychology;Surveillance
16 Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D. Abowd, and Munmun De Choudhury. 2023. Algorithmic Power or Punishment: Information Worker Perspectives on Passive Sensing Enabled AI Phenotyping of Performance and Wellbeing. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI ’23). Association for Computing Machinery, New York, NY, USA, Article 246, 1–17. https://doi.org/10.1145/3544548.3581376
17 Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L. Kun, Hagit Ben-Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation, In Proc. CHI 2024.
18 Toshali Goel, Orit Shaer, Catherine Delcourt, Quan Gu, and Angel Cooper. 2023. Preparing Future Designers for Human-AI Collaboration in Persona Creation. In Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK ’23). Association for Computing Machinery, New York, NY, USA, Article 4, 1–14. https://doi.org/10.1145/3596671.3598574
19 Markus Funk, Sven Mayer, and Albrecht Schmidt. 2015. Using In-Situ Projection to Support Cognitively Impaired Workers at the Workplace. In Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS ’15). Association for Computing Machinery, New York, NY, USA, 185–192. https://doi.org/10.1145/2700648.2809853
20 Don Samitha Elvitigala, Yunfan Wang, Yongquan Hu, and Aaron J Quigley. 2023. RadarFoot: Fine-grain Ground Surface Context Awareness for Smart Shoes. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23). Association for Computing Machinery, New York, NY, USA, Article 87, 1–13. https://doi.org/10.1145/3586183.3606738
21 Yongquan Hu, Hui-Shyong Yeo, Mingyue Yuan, Haoran Fan, Don Samitha Elvitigala, Wen Hu, and Aaron Quigley. 2023. MicroCam: Leveraging Smartphone Microscope Camera for Context-Aware Contact Surface Sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 7, 3, Article 98 (September 2023), 28 pages. https://doi.org/10.1145/3610921
22 Anne-Marie Mann, Uta Hinrichs, Janet C. Read, and Aaron Quigley. 2016. Facilitator, Functionary, Friend or Foe? Studying the Role of iPads within Learning Activities Across a

School Year. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 1833–1845. https://doi.org/10.1145/2858036.2858251

23 Wendy Martinez, Johann Benerradi, Serena Midha, Horia A. Maior, and Max L. Wilson. 2022. Understanding the Ethical Concerns for Neurotechnology in the Future of Work. In Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '22). Association for Computing Machinery, New York, NY, USA, Article 17, 1–19. https://doi.org/10.1145/3533406.3533423

24 Takumi Ito, Naomi Yamashita, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, Ge Gao, Jack Jamieson, and Kentaro Inui. 2023. Use of an AI-powered Rewriting Support Software in Context with Other Tools: A Study of Non-Native English Speakers. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 45, 1–13. https://doi.org/10.1145/3586183.3606810

25 Shadan Sadeghian and Marc Hassenzahl. 2022. The "Artificial" Colleague: Evaluation of Work Satisfaction in Collaboration with Non-human Coworkers. In 27th International Conference on Intelligent User Interfaces (IUI '22). Association for Computing Machinery, New York, NY, USA, 27–35. https://doi.org/10.1145/3490099.3511128

26 Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 4, Article 170 (December 2018), 22 pages. https://doi.org/10.1145/3287048

27 Unsoo Ha, Sohrab Madani, and Fadel Adib. 2021. WiStress: Contactless Stress Monitoring Using Wireless Signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 3, Article 103 (Sept 2021), 37 pages. https://doi.org/10.1145/3478121

11.5 Remote meetings

28 Alberta A. Ansah, Yilun Xing, Amudha Varshini Kamaraj, Diana Tosca, Linda Boyle, Shamsi Iqbal, Andrew L. Kun, John D. Lee, Michel Pahud, and Orit Shaer. 2022. "I need to respond to this" – Contributions to group creativity in remote meetings with distractions. In Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '22). Association for Computing Machinery, New York, NY, USA, Article 7, 1–12. https://doi.org/10.1145/3533406.3533411

29 B. Adriel Aseniero, M. Constantinides, S. Joglekar, K. Zhou and D. Quercia, "MeetCues: Supporting Online Meetings Experience," 2020 IEEE Visualization Conference (VIS), Salt Lake City, UT, USA, 2020, pp. 236-240, doi: 10.1109/VIS47514.2020.00054. keywords: Visualization;Conferences;Ecosystems;Tools;Real-time systems;Visualization for Meetings;Engagement;Awareness;Reflection

30 Hancheng Cao, Chia-Jung Lee, Shamsi Iqbal, Mary Czerwinski, Priscilla N Y Wong, Sean Rintel, Brent Hecht, Jaime Teevan, and Longqi Yang. 2021. Large Scale Analysis of Multitasking Behavior During Remote Meetings. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 448, 1–13. https://doi.org/10.1145/3411764.3445243

31 Yang, F., Heeman, P. A., and Kun, A. L. (2011). An investigation of interruptions and resumptions in multi-tasking dialogues. Computational Linguistics, 37(1), 75-104.

11.6 Healthcare

32 Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. Proc. ACM Hum.-Comput. Interact. 4, CSCW1, Article 31 (May 2020), 27 pages. https://doi.org/10.1145/3392836

33 Azra Ismail, Divy Thakkar, Neha Madhiwalla, and Neha Kumar. 2023. Public Health Calls for/with AI: An Ethnographic Perspective. Proc. ACM Hum.-Comput. Interact. 7, CSCW2, Article 354 (October 2023), 26 pages. https://doi.org/10.1145/3610203

34 Azra Ismail and Neha Kumar. 2021. AI in Global Health: The View from the Front Lines. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 598, 1–21. https://doi.org/10.1145/3411764.3445130

35 The AI Revolution in Medicine: GPT-4 and Beyond Paperback. 2023. by Peter Lee, Carey Goldberg, Isaac Kohane

36 Karthik S. Bhat, Neha Kumar, Karthik Shamanna, Nipun Kwatra, and Mohit Jain. 2023. Towards Intermediated Workflows for Hybrid Telemedicine. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 347, 1–17. https://doi.org/10.1145/3544548.3580653

37 Siddhartha Gairola, Murtuza Bohra, Nadeem Shaheer, Navya Jayaprakash, Pallavi Joshi, Anand Balasubramaniam, Kaushik Murali, Nipun Kwatra, and Mohit Jain. 2022. SmartKC: Smartphone-based Corneal Topographer for Keratoconus Detection. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 4, Article 155 (Dec 2021), 27 pages. https://doi.org/10.1145/3494982

38 Ayanna Seals, Monsurat Olaosebikan, Jennifer Otiono, J, Orit Shaer, Oded Nov (2021). Effects of Self-Focused Augmented Reality on Health Perceptions During the COVID-19 Pandemic: A Between-Subject Web-Based Experiment. Journal of medical Internet research, 10.2196/26963

11.7 Gig work

39 Juan Carlos Alvarez de la Vega, Marta E. Cecchinato, and John Rooksby. 2022. Design Opportunities for Freelancing Platforms: Online Freelancers' Views on a Worker-Centred Design Fiction. In Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '22). Association for Computing Machinery, New York, NY, USA, Article 12, 1–19. https://doi.org/10.1145/3533406.3533410

40 Juan Carlos Alvarez de la Vega, Marta E. Cecchinato, and John Rooksby. 2021. "Why lose control?" A Study of Freelancers' Experiences with Gig Economy Platforms. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 455, 1–14. https://doi.org/10.1145/3411764.3445305

Participants

- Larbi Abdenebaoui
OFFIS – Oldenburg, DE
- Susanne Boll
Universität Oldenburg, DE
- Duncan Brumby
University College London, GB
- Marta Cecchinato
University of Northumbria – Newcastle upon Tyne, GB
- Marios Constantinides
Nokia Bell Labs – Cambridge, GB
- Anna Cox
University College London, GB
- Mohit Jain
Microsoft Research India – Bangalore, IN
- Christian P. Janssen
Utrecht University, NL
- Naveena Karusala
Harvard University – Allston, US
- Neha Kumar
Georgia Institute of Technology – Atlanta, US
- Andrew L. Kun
University of New Hampshire – Durham, US
- Sven Mayer
LMU München, DE
- Phillippe Palanque
Paul Sabatier University – Toulouse, FR
- Bastian Pfleging
TU Bergakademie Freiberg, DE
- Aaron Quigley
CSIRO – Eveleigh, AU
- Michal Rinott
SHENKAR – Engineering. Design. Art – Ramat-Gan, IL
- Shadan Sadeghian
Universität Siegen, DE
- Stefan Schneegass
Universität Duisburg-Essen, DE
- Orit Shaer
Wellesley College, US
- Erin T. Solovey
Worcester Polytechnic Institute, US
- Tim C. Stratmann
OFFIS – Oldenburg, DE
- Dakuo Wang
Northeastern University – Boston, US
- Max L. Wilson
University of Nottingham, GB
- Naomi Yamashita
NTT – Kyoto, JP

Report from Dagstuhl Seminar 23461

Space and Artificial Intelligence

Sašo Džeroski*[1], Holger H. Hoos*[2], Bertrand Le Saux*[3], Leendert van der Torre*[4], and Ana Kostovska†[5]

1 **Jožef Stefan Institute – Ljubljana, SI.** saso.dzeroski@ijs.si
2 **RWTH Aachen, DE.** hh@aim.rwth-aachen.de
3 **ESA – Frascati, IT.** bertrand.le.saux@esa.int
4 **University of Luxembourg, LU.** leon.vandertorre@uni.lu
5 **Jožef Stefan Institute – Ljubljana, SI.** ana.kostovska@ijs.si

Abstract

This report documents the program and the outcomes of the Dagstuhl Seminar 23461 "Space and Artificial Intelligence". The seminar was interdisciplinary, situated at the intersection of research on AI / computer science and space research. Since each of these is a very wide field on its own, we focussed on a selection of topics from each of the two and their intersections.

On the artificial intelligence side, we focused on data-driven AI, which makes use of data in order to produce intelligent behaviour and notably includes machine learning approaches. We also considered knowledge-based AI, which is focussed on the explicit formalisation of human knowledge and its use for tasks such as reasoning, planning, and scheduling. On the space research side, we considered the two major branches of space operations (SO) and Earth observation (EO).

The seminar brought together a diverse set of players, including researchers from academia, on one hand, and practitioners from space agencies (ESA, NASA) and industry, on the other hand. The seminar included plenary talks and parallel group discussions. Through the plenary talks, we obtained insight into the state-of-the-art in the different areas of AI research and space research, and especially in their intersections. Through the parallel group discussions, we identified obstacles and challenges to further progress and charted directions for further work.

Seminar November 12–17, 2023 – https://www.dagstuhl.de/23461

2012 ACM Subject Classification Computing methodologies → Computer vision; Computing methodologies → Knowledge representation and reasoning; Computing methodologies → Planning and scheduling; Computing methodologies → Learning paradigms; Computing methodologies → Learning settings; Computing methodologies → Machine learning approaches; Computing methodologies → Modeling and simulation; Applied computing → Aerospace; Applied computing → Astronomy; Applied computing → Engineering; Applied computing → Earth and atmospheric sciences

Keywords and phrases Artificial Intelligence, Machine Learning, Data-based AI, Knowledge-based AI, Deep Learning, Foundation Models, Explainable Artificial Intelligence, Space Research, Space Operations, Earth Observation

Digital Object Identifier 10.4230/DagRep.13.11.72

* Editor / Organizer
† Editorial Assistant / Collector

1 Executive Summary

Sašo Džeroski (Jožef Stefan Institute – Ljubljana, SI, saso.dzeroski@ijs.si)
Holger H. Hoos (RWTH Aachen, DE, hh@aim.rwth-aachen.de)
Bertrand Le Saux (ESA – Frascati, IT, bertrand.le.saux@esa.int)
Leendert van der Torre(University of Luxembourg, LU, leon.vandertorre@uni.lu)

Scope of the Seminar

Our interdisciplinary seminar on *Space and Artificial Intelligence* was situated at the intersection of research on AI / computer science and space research. Since each of these is a very wide field on its own, below we give a broad outline of each of the two. We focus on the aspects that were topics of discussion at our Seminar.

Artificial intelligence studies computer systems that behave similarly to humans, in a way that the resulting behaviour would be considered intelligent if exhibited by humans. The field of AI thus focusses on the design and analysis of algorithms and systems that can replicate, support or surpass human perceptual, linguistic, and reasoning processes; learn, draw conclusions, and make predictions based on large or small quantities of data; replicate or enhance human perception; support humans in diagnosis, planning, scheduling, resource allocation, and decision making; and cooperate physically and intellectually with humans and other AI systems (`https://claire-ai.org/what-is-ai/`). All these topics are relevant for space research.

At a high level, AI can be categorised into three (non-exclusive) categories:

Data-driven AI makes use of data in order to produce intelligent behaviour; it prominently encompasses machine learning, data mining, and pattern recognition approaches and is often referred to simply as machine learning. In this area, methods based on neural networks have been particularly successful and, as a result, become a major focus of attention for the last decade, but many other approaches exist and continue to be used with considerable impact, including support vector machines and random forest models for supervised learning, and various types of clustering methods for unsupervised learning.

Knowledge-based AI is focussed on the explicit formalisation of human knowledge and its use for tasks such as reasoning, planning, and scheduling. Although knowledge-based AI is currently somewhat less prominent than data-driven AI, it has important and impactful uses, e.g., in ensuring the correctness of computer hard- and software, and in solving a broad range of real-world industrial optimisation problems. Many AI experts now believe that combinations of data-driven and knowledge-based methods are likely to provide the basis for next-generation trustworthy AI systems. In this context, explainable AI (XAI), where the results of AI solutions can be understood by humans, is gaining importance.

Embodied AI concerns the design and study of AI systems that interact directly with the physical world. This area is also known as robotics and has very important applications in an increasingly broad range of application sectors, including manufacturing, medicine, and agriculture. Interaction with the physical environment (including other robots and humans) poses unique challenges, e.g., in terms of safety, robustness, and real-time requirements. Most experts in the field of robotics make use of knowledge-based and data-driven approaches, in addition to specialised methods for dealing with the previously mentioned challenges.

Legal and ethical aspects of AI have also started to attract attention recently. The legal part includes laws that regulate the use and development of artificial intelligence. The ethical part is concerned with the moral behavior of humans as they design, make, use, and handle artificially intelligent systems, but also with the moral behavior of machines (machine ethics).

The different forms of AI can be applied to a variety of problems in space-related research, of which here we highlight two major branches:

Space Operations (SO) are concerned with all aspects of operating spacecraft, including the planning, implementing, and operating of all (also ground segment) systems required for reliable and efficient spaceflight missions. This includes all relevant mission operations, ground infrastructure, flight dynamics, mission planning, communications, and data acquisition functions. Large amounts of data about space operations are collected and can be utilized by ML / data-driven AI to address challenges that include autonomous spacecraft route planning, spacecraft anomaly detection, and optimal spacecraft operations.

Earth Observation (EO) is a major instrument for monitoring our planet, its land and ocean processes, and their dynamics. A large number of spacecraft carrying a broad range of instruments generate a wide variety of sensor data (active / passive) of many resolutions: With these data now accessible to researchers and agencies, as well as the general public, a final barrier remains the need to convert the enormous quantities of raw EO data (generated on a daily basis) into valuable information for making decisions and taking concrete actions, e.g., towards achieving the Sustainable Development Goals. Needless to say, the potential for applying AI and ML in this context is almost unlimited.

Many other space-related AI applications can be conceived, typically related to the use of ML for the analysis of data collected during specific space missions. These include, e.g., modeling and forecasting space weather, mapping planet surfaces, galaxy profiling, identifying exoplanets and their environment, as well as analyzing astro-biology data. Several of these belong to astronomy and concern data collected via astronomical observatories in orbit.

Seminar topics

The seminar covered many different aspects of Artificial Intelligence for space and touched upon a wide variety of topics. However, it focussed specifically on the following four topics – all of which are currently actively researched – structured along two dimensions (AI approaches and Space applications):

Data-driven AI, e.g., machine learning, for space. The first topic of the seminar addressed machine learning methods for the analysis of the ever larger quantities of data resulting from space related research and exploration, their current state-of-the-art, and directions for further development.

Knowledge-driven AI, e.g., explainable AI, for space. The second topic of the seminar was concerned with methods and techniques from knowledge representation and reasoning, and explainable AI, their current state-of-the-art, and directions for further development.

Space Operations applications of AI. The third topic of the seminar concerned various aspects of operating spacecraft and managing missions, the potential applications of AI in this area, and the challenges they pose for Artificial Intelligence methods.

Earth Observation applications of AI. The fourth topic of the seminar concerned various aspects of applying AI to Earth observation data, the vast variety of potential applications of AI in this area, and the challenges they pose for Artificial Intelligence methods.

Note that the topics along the space applications dimension interact strongly with the AI approaches dimension. For example, space operations applications, such as estimating the current and predicting the future states of spacecraft, have a strong temporal dimension requiring the use of data stream mining approaches from AI. On one hand, this poses challenges to address in the development of novel AI methods. On the other hand, this can provide excellent benchmarking opportunities for the evaluation of AI methods.

The above four topics were the focus of the seminar. Given the interests of the participants of the seminar, we also considered a few additional topics (to a lesser extent). These included, for example, legal, ethical, and social aspects of Space AI.

Structure of the seminar

The structure of our seminar was standard for Dagstuhl. We started with an introduction round on Monday morning. The majority of the time was taken by plenary talks and parallel discussions in working groups: There were two of the latter, one on Tuesday morning and one on Friday morning. The social event on Tursday afternoon included a visit to the Völklingen Ironworks UNESCO industrial heritage site and a dinner.

Plenary talks. Given the highly interdisciplinary nature of the seminar, participants from one discipline needed to be brought up to speed with the state of the art in the other relevant disciplines. Some of the talks were thus of an overview or tutorial nature. Examples of such talks are "Introduction to Space Operations" and "Introduction to Explainable Artificial Intelligence". Other talks were more specific, addressing particular AI methods or classes thereof or particular (areas of) AI applications in space research.

The plenary talks can be clustered into four different groups

- Plenary talks on machine learning,
- Plenary talks on explainable AI,
- Plenary talks on earth observation, and
- Plenary talks on space operations.

Parallel discussion in working groups. A substantial part of the seminar time was split into structured small-group work sessions. The aim of the structured work sessions was to address the focal topics of the seminar that were most interesting for the participants. The participants could more effectively share knowledge and experiences from their own areas of expertise in the smaller working groups. The highlights of these structured small-group sessions were presented to the seminar as a whole.

The parallel discussions in working groups on Tuesday morning addressed the following topics:

- Sustainable development goals and AI for good,
- AutoML and benchmarks,
- On-board and frugal AI, and
- Responsible AI.

The parallel discussions in working groups on Friday morning all addressed the same topic of challenges in AI & space and future research directions.

Outcomes of the seminar

The seminar brought together a diverse set of players. These included researchers from academia, on one hand, and practitioners from space agencies (ESA, NASA) and industry, on the other hand. It covered a broad range of aspects relevant for the further development of the field.

The major outcomes of the seminar are as follows:

1. It gave researchers from the different contributing disciplines an integrated overview of current research in the area of artificial intelligence for space.
2. It reinforced the communication channels for researchers tackling challenges in space applications using AI, including both data driven and knowledge-driven approaches to AI, such as machine learning and explainable AI, thereby bridging the divide between computer science and space research.
3. It defined the landscape of potential applications of artificial intelligence in space, in particular in the areas of Space Operations and Earth Observation.
4. It identified the central research questions and challenges for artificial intelligence approaches that need to be resolved for successful use of AI in space applications.
5. It put forward some strategies for designing artificial intelligence tools for space applications and for developing benchmarking suites for evaluating such approaches.

2 Table of Contents

3 Plenary talks: Machine Learning

3.1 Semi-supervised and multi-label classification of remotely sensed images

Sašo Džeroski (Jožef Stefan Institute – Ljubljana, SI, saso.dzeroski@ijs.si)

Joint work of Marjan Stoimchev, Jurica Levatić, Dragi Kocev, Michelangelo Ceci, Sašo Džeroski
Main reference Marjan Stoimchev, Dragi Kocev, Sašo Džeroski: "Deep Network Architectures as Feature Extractors for Multi-Label Classification of Remote Sensing Images", Remote. Sens., Vol. 15(2), p. 538, 2023.
URL https://doi.org//10.3390/RS15020538

The talk will discuss recent work on semi-supervised (SS) [4] and multi-label classification (MLC) of remotely sensed images (RSI) [1]. For MLC, we employ deep neural networks (DNNs), either as feature extractors for predictive clustering trees (PCTs) and ensembles thereof, or in an end-to-end manner. In the former case, we leverage the existing capabilities of semi-supervised PCTs and ensembles: explainability of single tree models and state-of-the-art predictive performance of random forest ensembles [3, 2]. Furthermore, the parametrization of the amount of supervision in PCTs allows us to build supervised, semi-supervised, or unsupervised models, depending on the demands of the dataset at hand. This provides a safety mechanism enabling the semi-supervised models to consistently perform better or as good as their supervised counterparts.

We also develop end-to-end semi-supervised DNNs for multi-label and multi-class classification of remotely sensed images. This method mimics the mechanism of semi-supervised PCTs that have been proven to work well. We introduce a novel loss function that combines classification loss (computed on labeled data) and reconstruction loss (computed on both labeled and unlabeled data) with a weight parameter that enables the same aforementioned "safety mechanism".

The capabilities of PCTs and ensembles of PCTs (e.g. Random Forests) enable us to perform hierarchical MLC of remotely sensed images – a novel formulation of the classification task in this field. To this end, we exploit the intrinsic label hierarchies of the BigEarthNet dataset and explore the effects different label hierarchies and their different handling have on predictive performance.

References

1 Stoimchev, Marjan; Kocev, Dragi and Džeroski, Sašo, *Deep Network Architectures as Feature Extractors for Multi-Label Classification of Remote Sensing Images*. Remote Sensing, 15(2), 538, 2023.
2 Levatić, Jurica; Kocev, Dragi; Ceci, Michelangelo and Džeroski, Sašo, *Semi-supervised trees for multi-target regression*. Information Sciences, 450, 109–127, 2018.
3 Levatić, Jurica; Ceci, Michelangelo; Kocev, Dragi and Džeroski, Sašo, *Semi-supervised classification trees*. Journal of Intelligent Information Systems, 49, 461–486, 2017.
4 Levatić, Jurica; Ceci, Michelangelo; Kocev, Dragi and Džeroski, Sašo, *Semi-supervised Predictive Clustering Trees for (Hierarchical) Multi-label Classification*. arXiv preprint arXiv:2207.09237, 2022.

3.2 Self-supervised Learning, Foundation Models, and ModelZoos

Damian Borth (University of St. Gallen, CH, damian.borth@unisg.ch)

Joint work of Damian Borth, Diyar Taskiran, Konstantin Schürholt, Boris Knyazev, Xavier Giró-i-Nieto
Main reference Konstantin Schürholt, Boris Knyazev, Xavier Giró-i-Nieto, Damian Borth: "Hyper-Representations as Generative Models: Sampling Unseen Neural Network Weights", in Proc. of the Advances in Neural Information Processing Systems, 2022.
URL https://openreview.net/forum?id=uyEYNg2HHFQ

Self-supervised learning allowed us to train large task agnostic backbones, which can be successfully finetuned for specialized downstream tasks with only little supervision. This opened the path towards the training of so-called foundation models, a family of task-agnostic representations potentially able to consume multiple modalities of inputs and able to not only encapsulate a wide range of known tasks but are also able to extend this range to new task with only few shots of example. One popular family of such foundation models are large-scale language models.

This talk will provide an overview of self-supervised learning, its pretext tasks, and proposed learning methods from the last years. It further introduces the idea of learning from populations of neural networks, so called model zoos and shows how task-agnostic representation from these model zoos – so called hyper-representations – can be learned. Finally, it demonstrates how these representations can be exploited for multiple discriminative and generative downstream tasks linking them to model diagnostic, inspection and model sampling, finetuning.

References

1 Honegger, D.; Schürholt, K.; Scheibenreif, L.; Borth, D., *Eurosat Model Zoo: A Dataset and Benchmark on Populations of Neural Networks and Its Sparsified Model Twins*. IGARSS, 2023.
2 Honegger, D.; Schürholt, K.; Borth, D., *Sparsified Model Zoo Twins: Investigating Populations of Sparsified Neural Network Models*. ICLR Workshop on Sparsity in Neural Networks (SNN), 2023.
3 Schürholt, K.; Knyazev, B.; Giró-i-Nieto, X.; Borth, D., *Hyper-Representations as Generative Models: Sampling Unseen Neural Network Weights*. Neural Information Processing Systems (NeurIPS), 2022.
4 Schürholt, K.; Taskiran, D.; Knyazev, B.; Giró-i-Nieto, X.; Borth, D., *Model Zoos: A Dataset of Diverse Populations of Neural Network Models*. Neural Information Processing Systems (NeurIPS), 2022.
5 Schürholt, K.; Knyazev, B.; Giró-i-Nieto, X.; Borth, D., *Hyper-Representations for Pre-Training and Transfer Learning*. ICML Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward, 2022.
6 Scheibenreif, L.; Hanna, J.; Mommert, M.; Borth, D., *Self-Supervised Vision Transformers for Land-Cover Segmentation and Classification*. CVPR EarthVision Workshop, 2022.
7 Schürholt, K.; Kostadinov, D.; Borth, D., *Hyper-Representations: Self-Supervised Representation Learning on Neural Network Weights for Model Characteristic Prediction*. Neural Information Processing Systems (NeurIPS), 2021.

3.3 Hybrid modelling: examples and challenges

Nuno Carvalhais (MPI for Biogeochemistry – Jena, DE, nuno.carvalhais@bgc-jena.mpg.de)

Joint work of Nuno Carvalhais, Shanning Bao, Rackhun Son, Christian Requena, Lazaro Alonso, Markus Reichstein
Main reference Rackhun Son, Tobias Stacke, Veronika Gayler, Julia E. M. S. Nabel, Reiner Schnur, Lazaro Alonso, Christian Requena-Mesa, Alexander J. Winkler, Stijn Hantson, Sönke Zaehle, Ulrich Weber, Nuno Carvalhais: "Integration of a Deep-Learning-Based Fire Model Into a Global Land Surface Model", Journal of Advances in Modeling Earth Systems, Vol. 16(1), p. e2023MS003710, 2024.
URL https://doi.org//10.1029/2023MS003710

Challenges in representing the spatial and temporal dynamics of carbon and water cycles in land ecosystems arise both from parametric and or epistemic uncertainties. Understanding and quantifying ecosystem responses to changes in climate and environmental conditions underpins the quantification of coupled climate-carbon cycle feedbacks, key for addressing today's Earth system challenges. The growing volume in Earth observation data delivers an unprecedented perspective for improving understanding as well as unprecedented challenges in traditional Earth system model development. Here, we propose two hybrid modelling approaches for maximizing the information content uptake in improving carbon cycle modelling leveraging EO and machine learning approaches.

On the one hand, we propose an end-to-end approach that learns the spatial variation in parameters controlling the daily to seasonal response of photosynthesis to climate and atmospheric CO2 as described by a light use efficiency (LUE) model. The LUE model parameters emerge from the outputs of a multi-layer perceptron fed by a set of features representing vegetation, soils and bioclimatic properties. The MLP learns from the minimization of mismatch between modelled and observed fluxes of carbon and water in eddy covariance sites. The cross-validation results show a robust comparison to observations, being close to calibration results, and the only parameter generalization approach robust to represent spatial and temporal patterns.

On the other hand, we explore the potential of infusing traditional process-based models with machine learning components to describe largely uncertain processes. We develop an MLP architecture standing on parallel long short-term memory components to represent the role of the atmosphere, land surface properties and anthropogenic features to predict burned area dynamics at global scales as observed from EO. Upon integrating the trained MLP within the process-based model we observe a stark contrast in model performance in comparison to the benchmark fire model. Reductions in performance in some regions of the globe in comparison to the EO-driven MLP are related to internal biases in process-based modelled state variables, reflecting the need to develop online training approaches.

Overall, while the predictive performance in hybrid modelling improves significantly from current baselines in process-based modelling approaches, we are challenged by features collinearity for attributing variability in parameters and global patterns in burned area dynamics. In a context of climate change, being able to appropriately attribute statistical and causal dependence in parameterizations and processes is key for advancing our understanding and quantification of Earth system dynamics.

References

1 Bao, Shanning; Alonso, Lazaro; Wang, Siyuan; Gensheimer,Johannes; De,Ranit and Carvalhais, Nuno, *Toward robust parameterizations in ecosystem-level photosynthesis models*. Journal of Advances in Modeling Earth Systems, 15, e2022MS003464, 2023.

2 Son, Rackhun; Stacke, Tobias; Gayler, Veronika; Nabel, Julia E. M. S.; Schnur, Reiner; Alonso, Lazaro; Requena-Mesa, Christian; Winkler, Alexander J.; Hantson, Stijn; Zaehle, Sönke; Weber, Ulrich; Carvalhais, Nuno, *Integration of a Deep-Learning-Based Fire Model Into a Global Land Surface Model*. Journal of Advances in Modeling Earth Systems, 2024.

3.4 Automated Machine Learning for SeaICE Charting

Jan N. van Rijn (Leiden Institute of Advanced Computer Science, NL, j.n.van.rijn@liacs.leidenuniv.nl)

Joint work of Jan N. van Rijn, Sven van Collenburg, Holger Hoos, Andreas Stokholm
Main reference Andreas Stokholm, Tore Wulf, Andrzej Kucik, Roberto Saldo, Jorgen Buus-Hinkler, Sine Munk Hvidegaard: "AI4SeaIce: Toward Solving Ambiguous SAR Textures in Convolutional Neural Networks for Automatic Sea Ice Concentration Charting", IEEE Trans. Geosci. Remote. Sens., Vol. 60, pp. 1–13, 2022.
URL https://doi.org//10.1109/TGRS.2022.3149323

In this presentation, I will discuss the work that we have carried out under the ESA-visiting professor program towards automated machine learning for sea ice charting. Sea ice charting is an important task for ships sailing across the North Pole, as the best sailing route depends on the location and type of sea ice. Where the charting process was originally carried out by professional charters who have access to SAR satellite data, artificial intelligence can now play an important role in supporting the charting expert.

In earlier work, Stokholm et al. [1] successfully trained a U-NET neural network on the SAR satellite images to predict the various classes of sea ice. Neural networks are highly sensitive to their hyperparameter settings, and properly tuning the hyperparameters can make the difference between mediocre performance and state-of-the-art performance. In this work, we set out to use automated machine learning (AutoML) to automate the hyperparameter tuning for this specific problem domain.

As this is an interdisciplinary audience, I will briefly cover the basics of AutoML, such as: why would we use AutoML, what is AutoML and what are basic algorithms in AutoML. Additionally, I will talk about how AutoML was used for this specific domain, and how it automated this crucial part of the data science loop.

References

1 Stokholm, Andreas and Wulf, Tore and Kucik, Andrzej and Saldo, Roberto and Buus-Hinkler, Jørgen and Hvidegaard, Sine Munk, *AI4SeaIce: Toward Solving Ambiguous SAR Textures in Convolutional Neural Networks for Automatic Sea Ice Concentration Charting*. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–13, 2022.

3.5 Automated Machine Learning for Spatio-temporal Datasets

Mitra Baratchi (Leiden University, NL, m.baratchi@liacs.leidenuniv.nl)

Joint work of Laurens Arp, Julia Wasals, Peter van Bodgem, Alistair Fracis, Suzanne Marselis, Michael Marszalek, Nuno Sa, Victor Neuteboom, Nguyen Dang, James Wheeler, Nicolas Longepe, Holger Hoos, Mitra Baratchi
Main reference Laurens Arp, Mitra Baratchi, Holger H. Hoos: "VPint: value propagation-based spatial interpolation", Data Min. Knowl. Discov., Vol. 36(5), pp. 1647–1678, 2022.
URL https://doi.org//10.1007/S10618-022-00843-2

Automated machine learning (AutoML) is a young research area aiming at making high-performance machine-learning techniques accessible to a broad set of domain users by identifying all design choices in creating a machine-learning model and addressing them automatically. In this talk, I provided a number of examples that show different opportunities provided by taking an Automated Machine Learning approach to address various AI problems based on Earth observations.

The first opportunity lies in making use of all existing solutions to create a search space composed of available algorithms for specific task. Taking an efficient search strategy to find models in this search space allows the configuration of customised models for each dataset automatically. I provided an example demonstrating this approach by showing how available deep learning algorithms for super-resolution can be used to create an AutoML system that configures deep learning models for super-resolution. The second opportunity demonstrates that adding new algorithmic solutions to the search space of AutoML systems can provide an opportunity to generate much stronger AutoML systems. As an example, I presented VPint [2], an interpolation algorithm and how it can be used to complement AuotML systems when used for cloud removal purposes.

Next, I presented an approach that allows to use the knowledge in a specific class of physical models called radiative transfer models to generate physics-aware machine learning pipelines. The first example [4] extends an existing AutoML system to create an ensemble of physics-driven and data-driven models. The second example [5] provides a framework to address the fundamental problem of ill-posedness of a class of physical models known as radiative transfer models.

This talk inspired a lively discussion on AutoML systems for Earth observation. Notably, the audience discussed the challenges of using different classes of physical models, for instance, radiative transfer models and dynamical system models. It also sparked discussions on different opportunities for performing cross-validation considering the spatial and temporal correlations.

References

1 Wąsala, Julia; Marselis, Suzanne; Arp, Laurens; Hoos, Holger; Longépé, Nicolas; Baratchi, Mitra, *AutoSR4EO: An AutoML approach to super-resolution for Earth observation images*. under review, 2023.

2 Arp, Laurens; Baratchi, Mitra; Hoos, Holger, *VPint: value propagation-based spatial interpolation*. Data Mining and Knowledge Discovery, 36(5), 1647-1678, 2022.

3 Arp, Laurens; Hoos, Holger; van Bodegom, Peter; Francis, Alice; Wheeler, Jennifer; Baratchi, Mitra, *Training-free thick cloud removal for Sentinel-2 imagery using value propagation interpolation*. Under review, 2023.

4 Neuteboom, Valerie; Baratchi, Mitra; van Bodegom, Peter; de Sá, Nuno C.; Marszalek, Martin, *AutoML for creating hybrid Earth science models*. Master's Thesis in Computer Science at Leiden Institute of Advanced Computer Science, Leiden University, 2021.

5 Arp, Laurens; van Bodegom, Peter; Dang, Nga; Francis, Alice; Hoos, Holger; Baratchi, Mitra, *Ill-posed inversion of physical models for parameter retrieval using diversity optimisation*. in preparation, 2023.

4 Plenary talks: Explainable Artificial Intelligence

4.1 Introduction to Explainable Artificial Intelligence

Yazan Mualla (University of Technology of Belfort-Montbéliard, FR, yazan.mualla@utbm.fr)

Joint work of Yazan Mualla, Amro Najjar
Main reference Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, Christophe Nicolle: "The quest of parsimonious XAI: A human-agent architecture for explanation formulation", Artif. Intell., Vol. 302, p. 103573, 2022.
URL https://doi.org//10.1016/J.ARTINT.2021.103573

Recently, eXplainable AI (XAI) gained momentum both in academia and industry to explain the results of black-box machine learning algorithms. A landscape of XAI branches along with strategies for developing explainable models are provided. Latest empirical studies have confirmed that explaining a system's behavior to human users fosters the latter's acceptance of the system. However, providing overwhelming or unnecessary information may also confuse the users and cause failure. For these reasons, parsimony has been outlined as one of the key features of XAI with parsimonious explanation defined as the simplest explanation that describes the situation adequately. Our work proposes HAExA, a human-agent explainability architecture to formulate parsimonious explanations for remote robots. This is particularly applicable to space since the communication with Earth has limited bandwidth and significant delay. Finally, some challenges, opportunities, and applications of XAI directed to different space stakeholders are presented.

References

1 Mualla, Yazan; Tchappi, Igor; Kampik, Timotheus; Najjar, Amro; Calvaresi, Davide; Abbas-Turki, Abdeljalil; Galland, Stéphane; Nicolle, Christophe, *The quest of parsimonious XAI: A human-agent architecture for explanation formulation*. Artificial Intelligence, Volume 302, 2022.
2 Gunning, David; Aha, David, *DARPA's explainable artificial intelligence (XAI) program*. AI magazine, 40(2), 44-58, 2019.
3 Miller, Tim, *Explanation in artificial intelligence: Insights from the social sciences*. Artificial intelligence, 267, 1-38, 2019.
4 Hoffman, Robert R.; Mueller, Shane T.; Klein, Gary; Litman, Jordan, *Metrics for explainable AI: Challenges and prospects*. arXiv preprint arXiv:1812.04608, 2018.

4.2 Causal inference for data-driven science

Jakob Runge (Institute of Data Science, German Aerospace Center – Jena, DE, Jakob.Runge@dlr.de)

Joint work of Jakob Runge, Andreas Gerhardus, Veronika Eyring, Gustau Camps-Valls
Main reference Jakob Runge, Andreas Gerhardus, Veronika Eyring, Gustau Camps-Valls: "Causal inference for time series". Nature Reviews Earth & Environment, 4(7), 487-505, 2023.
URL https://doi.org/10.1038/s43017-023-00431-y

Machine learning excels in learning associations and patterns from data and is increasingly adopted in natural-, life- and social sciences, as well as engineering[6]. However, many relevant research questions about such complex systems are inherently causal and machine learning alone is not designed to answer them [2]. At the same time there often exists ample theoretical and empirical knowledge in the application domains.

Causal inference provides the theoretical foundations to use data and qualitative domain knowledge to quantitatively answer these questions, complementing statistics and machine learning techniques [5, 3, 4]. Given the strong causal implications, the application of causal inference methods requires a thorough reasoning about the the appropriateness of the assumptions that can give rise to causal conclusions. Furthermore, causal methods still share the same challenges that affect the statistical and machine learning techniques that they employ, from finite sample issues to the problem of hyperparameter tuning and computational complexities.

A problem that is especially relevant in applications of causal inference concerns the broad language gap between the methodological and domain science communities. In this contribution [7], we explain the use of causal inference frameworks with a focus on the challenges of time series data and particular application scenarios, from process understanding to the evaluation and comparison of physical simulation models via causal methods. Integrating causal thinking into data-driven science will facilitate process understanding and more robust machine learning and statistical models for spatio-temporal problems in Earth sciences, allowing to tackle many open problems with relevant environmental, economic, and societal implications.

References

1 Runge, Jakob et al., *Inferring causation from time series in Earth system sciences*. Nature Communications, 10, 2553, 2019.
2 Runge, Jakob; Gerhardus, Andreas; Varando, Gherardo; Eyring, Veronika; Camps-Valls, Gustau, *Causal inference for time series*. Nature Reviews Earth & Environment, 4(7), 487-505.
3 Spirtes, Peter; Glymour, Clark; Scheines, Richard, *Causation, Prediction, and Search*. MIT Press, 2000.
4 Peters, Jonas; Janzing, Dominik; Schölkopf, Bernhard, *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.
5 Pearl, Judea, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.
6 Reichstein, Markus et al., *Deep learning and process understanding for data-driven Earth system science*. Nature, 566, 195–204, 2019.
7 Runge, Jakob; Gerhardus, Andreas; Varando, Gherardo; Eyring, Veronika; Camps-Valls, Gustau, *Causal inference for time series*. Nature Reviews Earth & Environment, 4(7), 487-505, 2023.

4.3 Causality is all you need

Gustau Camps-Valls (University of Valencia, ES, gustau.camps@uv.es)

This talk encapsulates a foundational exploration into causal inference, discovery, and effect estimation, offering a comprehensive 101 guide to the methods and techniques essential for understanding cause-and-effect relationships. This presentation navigated through theoretical and applied challenges, providing a holistic view of the complexities inherent in unravelling causation, especially for Earth and climate sciences. Some case studies served the purpose: from deciphering the drivers of migration to assessing the impact of humanitarian aid on food insecurity, as well as employing causal feature representation learning to unveil the influence of the El Niño-Southern Oscillation (ENSO) on vegetation greenness in Africa.

I also introduce the innovative causeme.net platform, a powerful tool for web-based causality analyses. This platform facilitates the exploration of causal relationships and is a practical resource for researchers and practitioners. The presentation concluded by teasing the integration possibilities between Large Language Models (LLMs) and causality studies. This forward-looking perspective hinted at the exciting potential for synergy between advanced language models and the nuanced understanding of causation, paving the way for future breakthroughs at the intersection of language processing and causal inference.

References

1 Cerdà, Jordi; Tarraga, Juan M.; Sevillano, Enrique; Munoz-Mari, Jorge; Piles, Maria; Camps-Valls, Gustau, *Understanding food insecurity in Africa through data-driven causal inference methods*. In World Climate Research Program, Kigali, Rwanda, December 2023. WRCP.

2 Tarraga, Juan M.; Sevillano, Enrique; Munoz-Mari, Jorge; Piles, Maria; Camps-Valls, Gustau, *Causal and explainable machine-learning models for hazard-induced displacement*. In World Climate Research Program, Kigali, Rwanda, December 2023. WRCP.

3 Ronco, Marco; Tárraga, Juan M.; Muñoz, Jorge; Piles, Maria; Sevillano Marco, Enrique; Wang, Qi; Miranda Espinosa, Maria T.; Ponserre, Sylvie; Camps-Valls, Gustau, *Exploring interactions between societal context and natural hazards on human population displacement*. Nature Communications, 2023.

4 Tárraga, Juan M.; Camps-Valls, Gustau; Sevillano-Marco, Enrique; Ronco, Marco; Piles, Maria; Muñoz-María, Jorge; Miranda, Maria T.; Cerdà, Jordi; Wang, Qi, *Causal discovery reveals contrasting mechanisms of drought-induced displacement in Somalia*. Global Environmental Change, 2024.

5 Varando, Gherardo; Fernández-Torres, Miguel-Angel; Camps-Valls, Gustau, *Learning Granger Causal Feature Representations*. In International Conference on Machine Learning – Workshop on Tackling Climate Change with Machine Learning. ICML, 2021.

6 Camps-Valls, Gustau; Gerhardus, Andreas; Ninad, Urmi; Varando, Gherardo; Martius, Georg; Balaguer-Ballester, Emili; Vinuesa, Ricardo; Diaz, Emiliano; Zanna, Laure; Runge, Jakob, *Discovering causal relations and equations from data*. Physics Reports, 1044, 1–68, 2023.

7 Runge, Jakob; Gerhardus, Andreas; Varando, Gherardo; Eyring, Veronika; Camps-Valls, Gustau, *Causal inference for time series*. Nature Reviews Earth & Environment, 4(7), 487-505, 2023.

5 Plenary talks: Space Operations

5.1 Introduction to Space Operations

Alessandro Donati (former AI and Operations Innovation Manager at ESA, retired, DE, alessandro.donati@lilium.de)

The talk introduces space operations fundamental concepts in brief. First, the phases in time. The operations preparation phase, includes the setting up and customization of the ground segment elements, made of hardware, software, procedures, the specialists training and the simulation campaign. Then the operations execution, split in LEOP, Commissioning phase, Routine phase and decommissioning. The second part addresses the two parallel chains of health caring of the spacecraft and the productive chain, made of planning, execution, payload data acquisition and dissemination. Both parallel chains incorporate a variety of tasks that can embed AI algorithms. The identified tasks are preparation, planning, execution, monitoring, forecasting, diagnostic, optimization.

5.2 Exploring Challenges and Innovations in the Space Domain: Curated Datasets, Optimization Problems, and Machine Learning Applications

Dario Izzo (European Space Agency's Advanced Concepts Team – Noordwijk, NL, dario.izzo@esa.int)

Joint work of Dario Izzo, Marcus Maertens, Gabriele Meoni, Thomas Uriot, Luis F. Simoes, Pablo Gomez, Dominik Dold, Simone D'Amico
Main reference Dario Izzo, Gabriele Meoni, Pablo Gomez, Domink Dold, Alexander Zoechbauer: "Selected Trends in Artificial Intelligence for Space Applications" in Artificial Intelligence for Space: AI4SPACE: Trends, Applications, and Perspectives (1st ed.). CRC Press, 2023
URL https://doi.org/10.1201/9781003366386

This presentation addresses several challenges in the space domain that have prompted the creation and dissemination of meticulously curated datasets and optimization problems, contributing to the broader academic community. An overview of significant challenges, including the Proba-V super resolution challenge [1], the data-driven "The OPS-SAT case"challenge [3], the collision avoidance challenge [4], and the pose estimation challenge [2], is provided. The talk delves into each challenge, offering brief insights into their objectives and methodologies, while also sharing select results achieved in these endeavors. Additionally, the application of Machine Learning (ML) inversion techniques within the domain of Geodesy for irregular solar system bodies is presented highlighting the utilization of ML methods to address challenges specific to Geodesy, showcasing results obtained through this innovative approach. The incorporation of ML in geodetic processes not only introduces a novel dimension to a traditional problem but also demonstrates its potential to yield meaningful insights on the internal structure of irregular bodies.

The talk aims to shed light on the interdisciplinary applications of ML techniques in the space domain, emphasizing the collaborative and innovative efforts that drive advancements in space-related research.

References

1 Märtens, Marcus; Izzo, Dario; Krzic, Andrej and Cox, Daniël, *Super-resolution of PROBA-V images using convolutional neural networks*. Astrodynamics, 3, 387–402, 2019.
2 Kisantal, Mate; Sharma, Sumant; Park, Tae Ha; Izzo, Dario; Märtens, Marcus and D'Amico, Simone, *Satellite pose estimation challenge: Dataset, competition design, and results*. IEEE Transactions on Aerospace and Electronic Systems, 56(5), 4083–4098, 2020.
3 Izzo, Dario; Meoni, Gabriele; Gómez, Pablo; Dold, Domink and Zoechbauer, Alexander, *Selected Trends in Artificial Intelligence for Space Applications*. arXiv preprint arXiv:2212.06662, 2022.
4 Uriot, Thomas; Izzo, Dario; Simões, Luís F; Abay, Rasit; Einecke, Nils; Rebhan, Sven; Martinez-Heras, Jose; Letizia, Francesca; Siminski, Jan and Merz, Klaus, *Spacecraft collision avoidance challenge: Design and results of a machine learning competition*. Astrodynamics, 6(2), 121–140, 2022.
5 Izzo, Dario and Gómez, Pablo, *Geodesy of irregular small bodies via neural density fields*. Communications Engineering, 1(1), 48, 2022.

5.3 Architecting a data-driven future in space

Dan Crichton (Jet Propulsion Laboratory, California Institute of Technology – Pasadena, CA, USA, daniel.j.crichton@jpl.nasa.gov)

Joint work of Dan Crichton, Steve Chien, Richard Doyle, Riley Duren, Thomas Huang, Lukas Mandrake, Ben Smith, Hui Su
Main reference Daniel J. Crichton, Chris A. Mattmann, John S. Hughes, Sean C. Kelly, Andrew F. Hart: "A Multidisciplinary, Model-Driven, Distributed Science Data System Architecture", pp. 117–143, Springer, 2011.
URL https://doi.org//10.1007/978-0-85729-439-5_5

JPL and NASA have achieved unprecedented scientific understanding using remote sensing to explore of our solar system, the mysteries of the universe, and our home planet, Earth. Significant technical progress in mission capabilities and remote sensing instruments has dramatically changed over JPL's history. Missions today generate immense volumes of data, challenging conventional methods for capturing, managing, analyzing, and deriving insights from this wealth of information. Further, computational constraints onboard, coupled with bandwidth limitations in being able to transfer data the ground, require new innovative approaches to optimizing science yield and mission. Areas such as mission planning, onboard and ground-based data and science processing, data management, and science analysis can all benefit from new approaches in data science, artificial intelligence, autonomy, and computing.

JPL has already made substantial progress in these domains. Examples include onboard planning to facilitate more autonomous operations, real-time detection of transient events on Mars' surface, machine learning algorithms capable of identifying and classifying features in imaging, and the development of massively scalable data repositories to enable data mining. Much of these advances have been built on pioneering work JPL performed in areas such as machine learning applied to optical astronomy in the 1990s for analyzing images captured in nightly sky surveys. These breakthroughs have allowed JPL to continue to respond to opportunities to bring new computing capabilities to support both space mission operations and science.

This presentation will discuss the progress, challenges, and opportunities in applying data science, AI, software, and computing to space observing systems. It will present use cases and examples of successful operational deployments. Finally, it will explore the integration of these capabilities and their criticality for advancing next generation data-driven space observing architectures, highlighting areas for future research to scale computing capabilities in space and on the ground.

References

1 Chien, S. and Sherwood, R. and Tran, D. and Cichy, B. and Rabideau, G. and Castano, R. and Davis, A. and Mandl, D. and Frye, S. and Trout, B. and Shulman, S. and Boyer, D., *Using Autonomy Flight Software to Improve Science Return on Earth Observing One*. Journal of Aerospace Computing, Information, and Communication, 2(4), 196-216, 2005.

2 Crichton, D. et al., *A Scalable Planetary Science Information Architecture for Big Science Data*. 2014 IEEE 10th International Conference on e-Science, Sao Paulo, Brazil, 196-203, 2014.

3 Duren RM and Thorpe AK and Foster KT and Rafiq T and Hopkins FM and Yadav V and Bue BD and Thompson DR and Conley S and Colombi NK and Frankenberg C and McCubbin IB and Eastwood ML and Falk M and Herner JD and Croes BE and Green RO and Miller CE, *California's methane super-emitters*. Nature, 575(7781), 180-184, 2019.

4 Su, Hui and Wu, Longtao and Jiang, Jonathan H. and Pai, Raksha and Liu, Alex and Zhai, Albert J. and Tavallali, Peyman and DeMaria, Mark, *Applying Satellite Observations of Tropical Cyclone Internal Structures to Rapid Intensification Forecast With Machine Learning*. Geophysical Research Letters, 47(17), Art. No. e2020GL089102, 2020.

5 W. M. Keck Institute for Space Studies (KISS), *Nebulae: Deep-Space Computing Clouds*. KISS Study, August 26-30, 2019.

5.4 Challenges in fielding AI in Space Operations

Simone Fratini, Jose Martinez Heras (Solenix Engineering GmbH – Darmstadt, DE, simone.fratini@solenix.ch, jose.martinez@solenix.ch)

AI is a game-changer that is gathering momentum in space activities, as a prominent building block of enabling technology for future missions. But as much has been done so far, still many challenges remain.

In our talk we discussed some open challenges that, from our experience, when tackled, can bring substantial benefits for AI in space operations.

First, we acknowledge that even if operators need anomaly detection, we are only able to offer novelty detection. In most cases, novelty detection is close enough to anomaly detection to be useful. From our experience in Space Operations we highlight the problem of false alarms which drive the counter intuitive preference for precision (in detriment of recall), as having false alarms will cause operators to stop looking at any novelty detection system. We also stress the importance of being able to detect first time anomalies (i.e., anomalies that nobody thought this could happen) as they have the biggest impact in space operations.

Regarding diagnostics, we discuss several attempts and their limitations. Smart filtering can reduce the number of telemetry parameters operators need to consider but it cannot tell cause from effect, and it often produces many results. With Dependency Finder we can learn the relationship between different parameters from data; however, while it is useful to gain understanding, it cannot be used for deriving causes for a particular anomaly as it is based on large amounts of data. With Explainable AI (i.e., SHAP) we can tell which features are more relevant to get predictions in a Machine Learning model; however, it is not a causal relationship but a predictive relationship.

In terms of fielding AI for space operations, 4 levels of support have been discussed: augmentation, assistance, decision automation and autonomy. The variety and heterogeneity of knowledge to be engineered to implement AI-based support in operation, as well as the diversity of tasks to be considered, suggest that an hybrid approach would probably be the best option, combining various AI approaches as learning, modeling, reasoning and interaction.

In terms of autonomy, it has been discussed how many autonomous capabilities we will need in the near future, where and to do what. 4 scenarios of increasing autonomous levels where discussed: Augmentation and Support (the AI suggest and enhance the human being), Reactivity and Adaptiveness (the AI can perceive and respond to changes), Proactiveness (the AI can initiate action to meet its objectives) and Autonomy (the AI collaborates with the human being, "peer-to-peer").

It was finally pointed out as XAI and AI qualification more in general will be driving factors and essential enabling factors for a successful fielding of AI technologies in operations.

References

1 Martinez-Heras, Jose and Donati, Alessandro, *Enhanced telemetry monitoring with novelty detection*. AI Magazine, 35(4), 37-46, 2014.
2 Martinez-Heras, Jose and Donati, Alessandro and Sousa, Bruno and Fischer, Jörg, *DrMUST-a data mining approach for anomaly investigation*. In SpaceOps 2012, p. 1275109, 2012.
3 Martinez-Heras, Jose and Lukas, Luke and Donati, Alessandro, *Dependency Finder: Surprising Relationships in Telemetry*. In 2018 SpaceOps Conference, p. 2696, 2018.
4 Fratini, S. and Policella, N. and Silva, R. et al., *On-board autonomy operations for OPS-SAT experiment*. Applied Intelligence, (2021). https://doi.org/10.1007/s10489-020-02158-5
5 Gao, Y. and Burroughes, G. and Ocón, J. and Fratini, S. and Policella, N. and Donati, A., *Mission Operations and Autonomy*. In Y. Gao (Ed.), Contemporary Planetary Robotics: An Approach Toward Autonomous Systems, Wiley-VCH, 2016. ISBN: 978-3-527-41325-6.
6 Cesta, A. and Cortellessa, G. and Fratini, S. and Oddi, A. and Bernardi, G., *Deploying Interactive Mission Planning Tools – Experiences and Lessons Learned*. Journal of Advanced Computational Intelligence and Intelligent Informatics, 15(8), 1149-1158, 2011.

6 Plenary talks: Earth Observation

6.1 Foundational Models for Earth Observation

Bertrand Le Saux (European Space Agency – Frascati, IT, Bertrand.Le.Saux@esa.int)

Over the past decade, Earth Observation (EO) has undergone a significant transformation thanks to the deep learning revolution and the increasing use of Artificial Intelligence (AI) to address various EO challenges. Although a standard framework for developing AI solutions to EO problems has been established, it still faces several challenges such as lack of labeled data, generative modeling, and integration of physics.

Foundational models offer a new perspective by integrating unsupervisedly learned knowledge in large models that can be adapted to various use-cases. The Phileo foundational model is presented as an example of desirable features, including global-scale training and the combination of pillar models with various pretext tasks. The European Space Agency has commissioned future European Foundational Models for EO and Society, as well as Climate, which will further advance this field.

6.2 How can the EO "revolution" benefit NWP and climate prediction?

Jonathan Bamber, (University of Bristol, GB, J.Bamber@bristol.ac.uk)

In the last decade there has been an exponential rise in New Space missions, many with an EO focus. There is limited coordination between national space agencies and commercial actors in the space sector while at the same time there is an urgent need to improve forecast skill, range and robustness in numerical weather prediction but also for longer term climate projections.

Conventional modelling approaches are reaching the limit of computational capability as well as hard limits in power consumption. How can ML methods be best used to improve forecast skill, computational efficiency and data fusion in a highly distributed data centre structure? Should the focus be on Open Source foundation models that can be a community tool or hybrid model approaches or DTE types of approach? And what are the risks of one or two Big Tech companies developing a monopoly in the field and pushing not for profit organisations such as ECMWF out of the market?

6.3 Planning satellite observations for global monitoring of physical parameters: Some research questions

Gauthier Picard (ONERA/DTIS, Université de Toulouse, FR, gauthier.picard@onera.fr)

Joint work of Jean-Loup Farges, Gauthier Picard, Cédric Pralet

Operating Earth observation satellite constellation raises many challenges for AI and Agent-based Approaches [1]. Here, we identify questions about planning observation tasks as to monitor physical parameters using taskable agile EO satellites. Notably,

1- For monitoring physical phenomena, each acquisition carried out by a satellite provides not only instantaneous information on the observed area, but also (a) information on neighboring areas due to spatial correlations and (b) information on the value of physical parameters in the future, due to temporal dependencies. Therefore, a difficulty lies in assessing the value of each observation, knowing that due to limited capacity, planning systems for satellite constellations must select observations from a set of candidate observations [2, 3]. 2- The satellites considered can be equipped with different sensors, and some satellites are even capable of carrying out observations in several modes (for example, "wide field observation" mode versus "targeted observation" mode). Managing this heterogeneity is also a challenge for evaluating the reward associated with each observation. For example, it is necessary to find a compromise between, on the one hand, continuously maintaining global knowledge on the value of physical parameters, and on the other hand, carrying out targeted observations on areas where phenomena have been detected. 3- The choice of a good observation strategy depends on the dynamics of the physical process being monitored. For example, monitoring deforestation does not require the same frequency of observation as monitoring illicit degassing from ships. Ideally, the automated planning system should be able to learn a good observation strategy from a global request to monitor a parameter, rather than waiting for basic observation requests formulated by users.

References

1 Picard, Gauthier; Caron, Clément; Farges, Jean-Loup; Guerra, Jonathan; Pralet, Cédric; Roussel, Stéphanie, *Autonomous Agents and Multiagent Systems Challenges in Earth Observation Satellite Constellations*. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '21), International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 39–44, 2021.

2 Squillaci, Samuel; Pralet, Cédric; Roussel, Stéphanie, *Scheduling Complex Observation Requests for a Constellation of Satellites: Large Neighborhood Search Approaches*. In Integration of Constraint Programming, Artificial Intelligence, and Operations Research: 20th International Conference, CPAIOR 2023, Springer-Verlag, Berlin, Heidelberg, 443–459, 2023.

3 El Haouari, Jihanne, *Earth Observation Satellite scheduling under weather uncertainties*. Master Thesis Report, Université Toulouse, ENAC, ONERA, 2022. Available online: `https://www.gauthier-picard.info/files/masterthesisJEH.pdf`.

6.4 Artificial Intelligence and Earth Observation for The Sustainable Development Goals

Claudio Persello (University of Twente, Faculty of Geo-information Science and Earth Observation – Enschede, NL, c.persello@utwente.nl)

Main reference C. Persello, Jan Dirk Wegner, Ronny Hansch, Devis Tuia, Pedram Ghamisi, M. Koeva, Gustau Camps-Valls: "Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities", pp. 172–200, IEEE, 2022.
URL https://doi.org//10.1109/MGRS.2021.3136100

The combination of Artificial Intelligence (AI) and Earth observation (EO) promises significant advances to support the United Nations' Sustainable Development Goals (SDGs). New developments and applications are already changing how humanity will face our planet's challenges. This talk provides an overview of the areas where AI and EO can contribute the most towards achieving the SDGs, discussing opportunities and open challenges. Research activities on AI methods for EO data are presented along with their applications toward monitoring the progress and achieving the SDGs. Case studies are presented to achieve zero hunger (SDG 2), create sustainable cities (SDG 11), deliver tenure security (multiple SDGs), and mitigate and adapt to climate change (SDG 13). Important societal, economic, and environmental implications are covered.

References

1 Kuffer, Monika; Wang, Jiong; Nagenborg, Michael; Pfeffer, Karin; Kohli, Divyani; Sliuzas, Richard; Persello, Claudio, *The scope of earth-observation to improve the consistency of the SDG slum indicator*. ISPRS International Journal of Geo-Information, 7(11), 428, 2018.

2 Persello, Claudio; Stein, Alfred, *Deep fully convolutional networks for the detection of informal settlements in VHR images*. IEEE Geoscience and Remote Sensing Letters, 14(12), 2325-2329, 2017.

3 Persello, Claudio; Tolpekin, Valentyn A.; Bergado, John R.; De By, Rolf A., *Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping*. Remote Sensing of Environment, 231, 111253, 2019.

4 Persello, Claudio; Grift, John; Fan, Xiang; Paris, Cristina; Hänsch, Ronny; Koeva, Mila; Nelson, A., *AI4SmallFarms: A Data Set for Crop Field Delineation in Southeast Asian Smallholder Farms*. IEEE Geoscience and Remote Sensing Letters, 2023.

5 Tareke, Berhanu; Koeva, Mila; Persello, Claudio, *Extracting Polygons of Visible Cadastral Boundaries Using Deep Learning*. In IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, 915-918, IEEE, July 2023.

6.5 In-Domain Self-Supervised Learning Improves Remote Sensing Image Scene Classification

Sašo Džeroski (Jožef Stefan Institute – Ljubljana, SI, saso.dzeroski@ijs.si)

Joint work of Ivica Dimitrovski, Ivan Kitanovski, Nikola Simidjievski, Dragi Kocev
Main reference Ivica Dimitrovski, Ivan Kitanovski, Nikola Simidjievski, Dragi Kocev: "In-Domain Self-Supervised Learning Improves Remote Sensing Image Scene Classification", IEEE Geoscience and Remote Sensing Letters, Vol. 21, p. 1–5, Institute of Electrical and Electronics Engineers (IEEE), 2024.
URL https://doi.org//10.1109/lgrs.2024.3352926

We investigate the utility of in-domain self-supervised pre-training of vision models in the analysis of remote sensing imagery. Self-supervised learning (SSL) has emerged as a promising approach for remote sensing image classification due to its ability to exploit large amounts of unlabeled data. Unlike traditional supervised learning, SSL aims to learn representations of data without the need for explicit labels. This is achieved by formulating auxiliary tasks that can be used for pre-training models before fine-tuning them on a given downstream task. A common approach in practice to SSL pre-training is utilizing standard pre-training datasets, such as ImageNet. While relevant, such a general approach can have a sub-optimal influence on the downstream performance of models, especially on tasks from challenging domains such as remote sensing. In this paper, we analyze the effectiveness of SSL pre-training by employing the iBOT framework coupled with Vision transformers trained on Million-AID, a large and unlabeled remote sensing dataset. We present a comprehensive study of different self-supervised pre-training strategies and evaluate their effect across 14 downstream datasets with diverse properties. Our results demonstrate that leveraging large in-domain datasets for self-supervised pre-training consistently leads to improved predictive downstream performance, compared to the standard approaches found in practice.

7 Parallel working group discussions on different topics

7.1 Working Group on SDG and AI4Good

Jonathan Bamber (University of Bristol, Technical University of Munich, DE)
Xiaoxiang Zhu (Technical University of Munich, DE)
Gustau Camps-Valls (University of Valencia, ES)
Žiga Kokalj (Research Centre of Slovenian Academy of Sciences and Arts – Ljubljana, SI)
Jose Martinez-Heras (Solenix, DE)
Claudio Persello (University of Twente – Enschede, NL)
Michelangelo Ceci (University of Bari, IT)
Dino Ienco (INRAE – MOntpellier, FR)
Sašo Džeroski (Jožef Stefan Institute – Ljubljana, SI)
Alessandro Donati (European Space Agency, DE)
Sylvain Lobry (Université Paris Cité, FR)
Nuno Carvalhais (Max Planck Institute for Biogeochemistry – Jena, DE)

7.1.1 Discussed Problems

The group discussed several topics related to the promises and the challenges of using Artificial Intelligence (AI) and Earth Observation (EO) to support the Sustainable Development Goals (SDGs). First, the group discussed the possible tasks in SDG using AI and EO. These include the i) identification of the deprivation areas (for example, identify slums, informal settlements and inadequate houses, ii) identification of the best areas for installing renewable energy plants, iii) delineation of smallholder farms and semi-automated extraction of cadastral boundaries, iv) study of the local impacts of changes in climate and extreme events. The group then identified challenges in SDG using AI and EO. These include i) data quality, uncertainty quantification and difficulty in identifying the classes, ii) transferability of the models from one place to another vs. locality of the models, iii) ethical and privacy issues that apply to EO data, especially for high-resolution images. The last main topic discussed by the group concerns the exploitation of the models. Specifically, the problems that emerged are: i) policy makers need to make decisions, but also track the impact of the implementation of such decisions; ii) the implementation of these decisions largely depends on different governments and it's outside the scope of the influence of the EO and AI experts.

7.1.2 Conclusions

The contribution of AI and EO for SDG was considered very relevant by the group in order to solve many relevant tasks discussed during the meeting. However, several issues arise, which require the consideration of additional technologies, data and techniques (for example, integrating additional data to compensate the low-quality data available or the absence of some data). Transferability is also not easy and requires taking into account the peculiarities of the specific places. In terms of the exploitation of findings and results, it is necessary to call for outreach activities to present scientific results in a convincing way to compel politicians to act. An important point is to shift from model-centric towards data and user-centric AI.

7.2 Working Group on AutoML and Benchmarks

Marjan Stoimchev (Jožef Stefan Institute – Ljubljana, SI)
Ana Kostovska (Jožef Stefan Institute – Ljubljana, SI)
Pančе Panov (Jožef Stefan Institute – Ljubljana, SI)
Jurica Levatic (Jožef Stefan Institute – Ljubljana, SI)
Mitra Baratchi (Leiden University, NL)
Jan N. van Rijn (Leiden Institute of Advanced Computer Science, NL)
Joaquin Vanschoren (Eindhoven University of Technology, NL)

This working group concentrated on the use of AutoML methods in Space Operations and Earth Observation, as well as the presence of benchmarks.

7.2.1 Discussed Problems

During the breakout session, a variety of challenges and potential strategies surrounding the evaluation and development of Space Operations (SpaceOps) and Earth Observation (EO) foundational models were discussed. The conversation opened with questions about the current methods used to evaluate these models and whether foundational models truly offer superior performance. This led to a broader discussion on the necessity of creating new benchmarks specifically designed for EO data, highlighting the potential to develop something akin to a meta-album for Earth observation. Such benchmarks could cover a wide range of tasks, including image segmentation, pixel-level classification, and the use of multi-spectral data, underscoring the value of both labeled and unlabeled data for training purposes.

The session also tackled the difficulties AI researchers face in accessing EO data, emphasizing the need for better data availability to advance the field. The discussion acknowledged the complexity of integrating data from diverse sources, including different satellites, resolutions, and types (visual vs. radar), and the challenges of domain transfer, class-incremental learning, and cross-sentinel data integration. These issues highlight the need for benchmarks that can accommodate a variety of data characteristics and learning tasks.

Operational and technical challenges, such as data storage, licensing, and the infrastructure needed to host and share data and models, were also identified as significant hurdles. In this context, the potential role of AutoML in enhancing EO model development was explored. AutoML could simplify the search for optimal model configurations, leverage pretrained models for better efficiency and transferability, and help in defining effective search spaces.

7.2.2 Conclusions

The session concluded with a strong consensus on the need for new benchmarks and datasets that accurately reflect the complexities of Earth observation tasks. Engaging with the space community, utilizing resources like the upcoming Anomaly Detection dataset from ESA, and leveraging platforms such as kelvins.esa.int were identified as crucial steps forward. Moreover, operational considerations such as addressing data licensing and storage, and creating an infrastructure for model sharing, were acknowledged as essential for the progress of the field. The development of a benchmark, encompassing the selection of data, tasks, and metrics, was highlighted as a key action point, alongside the formation of a project team dedicated to building this benchmark. This collaborative approach, including research into the application of AutoML for foundational models, aims to overcome the challenges discussed and advance the field of Earth observation.

7.3 Working Group on On-board and Frugal AI

Damian Borth (University of St. Gallen, CH)
Dan Crichton (Jet Propulsion Laboratory, California Institute of Technology – Pasadena, US)
Simone Fratini (Solenix Engineering GmbH – Darmstadt, DE)
Holger Hoos (RWTH Aachen, DE)
Dario Izzo (European Space Agency's Advanced Concepts Team – Noordwijk, NL)
Gauthier Picard (ONERA/DTIS, Université de Toulouse, FR)
Jakub Nalepa (Silesian University of Technology – Gliwice, PL)

7.3.1 Discussed Problems

The discussions revolved around the necessity for frugality in AI application within space systems, particularly emphasizing the efficient use of training data, computational efforts, and the size/efficiency of the trained models. This approach, aimed at optimizing performance with minimal resource expenditure, is critical in scenarios where AI models are trained on Earth and deployed for inference on board spacecraft or satellites. The debate extended to whether the focus should solely be on machine learning (ML) techniques or include other AI approaches such as planning and scheduling, which are vital for autonomous decision-making in space.

A significant issue identified was the current state of onboard AI compared to ground operations, which is considered unsatisfactory. Despite planned developments for AI accelerators and dedicated software, there is a pressing need for broader community support to enhance these technologies' capabilities. Onboard autonomy in space requires a blend of model-based decision-making for task allocation and resource optimization, and learning for various functions including science inference and image analytics. AutoML and AutoAI emerge as potential solutions to reduce computational demands by automating algorithm selection and configuration, as well as performance prediction.

Another discussed problem was the need for multi-objective AutoML to balance performance with resource usage effectively, considering computation, memory, bandwidth, and response time. Furthermore, robustness against adversarial attacks and security concerns for open-source systems were highlighted, alongside the need for technology controlled by trusted entities to mitigate reliance on commercial enterprises.

7.3.2 Conclusions

The challenges in implementing frugal AI and autonomy in space are manifold. First, developing AI that can efficiently operate with limited resources on board, closing the sense-plan-act loop, remains a daunting task. This includes the integration of model-based planning, scheduling, and learning mechanisms that are capable of adapting to the dynamic space environment.

Second, the adoption of AutoML and other automated AI approaches requires advancements in multi-objective optimization to navigate the trade-offs between performance and resource consumption. This is particularly relevant for tasks like Earth Observation (EO), where satellites need to make autonomous decisions based on real-time data, such as cloud coverage.

Third, ensuring the robustness of onboard AI systems against adversarial attacks and addressing security vulnerabilities in open-source software are critical for maintaining the integrity and reliability of space missions. This is compounded by the challenge of developing and deploying AI technologies that remain under the control of trusted entities, avoiding over-reliance on commercial solutions.

Lastly, the computational complexity of combining data-driven and other AI methodologies for effective problem-solving in space poses a significant challenge. This includes optimizing task allocations and resource management in agile EO satellites, which require sophisticated planning and machine learning strategies to adapt to changing conditions and priorities.

Addressing these challenges necessitates a collaborative effort from the global research community, focusing on the development of advanced AI technologies that are efficient, secure, and capable of autonomous operation in the demanding conditions of space.

7.4 Working Group on Responsible AI

Leendert von der Torre (University of Luxembourg, LU)
George Anthony Long (Legal Parallax – Fountain Hills, US)
Yazan Mualla (University of Technology of Belfort-Montbéliard, FR)
Alexandru Tantar (Luxembourg Institute of Science and Technology, LU)
Bertrand Le Saux (European Space Agency – Frascati, IT)

7.4.1 Discussed Problems

The discussions addressed several critical issues regarding the regulation and application of artificial intelligence (AI) from a legal perspective. A key point was the unique nature of AI as a commodity, which necessitates a different regulatory approach than that applied to other powerful technologies, such as nuclear power. This distinction raises questions about how to appropriately regulate AI to ensure safety and accountability without stifling innovation.

Another significant issue was the differentiation between humans and AI-enhanced humans, particularly in legal contexts. Traditional punitive measures, like imprisonment, are not applicable to AI systems, highlighting the need for a responsible human principal behind AI operations. This situation parallels the legal treatment of companies but introduces complexities due to the varied intelligence levels of AI systems, ranging from highly intelligent to rudimentary.

The application of AI in space presents unique challenges, distinct from those on Earth. Privacy concerns, prevalent on Earth, are less relevant in space, where issues such as sovereignty, resource utilization, and the ethical implications of AI in scenarios without human intervention come to the forefront. These challenges underscore the need for international legal harmonization and collaboration in space activities.

7.4.2 Conclusions

The discussion highlights the critical need for a refined and tailored approach to the regulation of AI. The concept of "trustworthy AI," encompassing ethics, responsibility, and explainability, is identified as crucial for guiding future regulatory frameworks. However, implementing these principles faces significant challenges, both on Earth and in space.

On Earth, the focus is on creating unbiased datasets and ensuring fairness within the bounds of domestic laws. In space, the challenges are amplified, with concerns about resource management, ethical decision-making in critical scenarios, and the need for AI to operate with limited human intervention. The EU act "7 key requirements that AI systems should meet in order to be deemed trustworthy" highlight the importance of human oversight, safety, privacy, transparency, and accountability, which are particularly pertinent in the context of space.

Challenges arise in achieving a global consensus on AI regulation, especially for space activities. Despite these challenges, there is a shared commitment to developing AI that benefits humanity, with an emphasis on responsible innovation and the pursuit of harmonized standards on a global scale.

8 Parallel working group discussions on challenges in AI & space and future research directions

On Friday morning, the participants split into four discussion groups. All groups discussed the same general topic of challenges for AI and space, as well as future research directions. Summaries of the discussions, per working group, are given below.

8.1 Working Group 1

Damian Borth (University of St. Gallen, CH)
Dan Crichton (Jet Propulsion Laboratory, California Institute of Technology – Pasadena, US)
Alessandro Donati (European Space Agency, DE)
George Anthony Long (Legal Parallax – Fountain Hills, US)
Evridiki Ntagiou (ESA / ESOC – Darmstadt, DE)
Claudio Persello (University of Twente – Enschede, NL)
Joaquin Vanschoren (Eindhoven University of Technology, NL)
Žiga Kokalj (Research Centre of Slovenian Academy of Sciences and Arts – Ljubljana, SI)

From the discussion of this group, a number of short term and long-term opportunities were identified. Short-term opportunities include:

- Continuous engagement with users
- Increased collaboration between agencies and research facilities
- Make agencies (ESA, NASA, etc) visible to the research community
- Connect ESA and NASA to agree on some agreements to be competitive world-wide

Long-term opportunities include:
- International agreements on common approaches, standardization, and shared capabilities in AI/ML
- Space-based Autonomy for longer-term flights, system level control and onboard decisions
- Use of AI for space mining

8.2 Working Group 2

Michelangelo Ceci (University of Bari, IT)
Michai Datcu (University Politehnica of Bucharest, RO)
Simone Fratini (Solenix Engineering GmbH – Darmstadt, DE)
Dario Izzo (European Space Agency's Advanced Concepts Team – Noordwijk, NL)
Marjan Stoimchev (Jožef Stefan Institute – Ljubljana, SI)

The working group 2 discussed the following points:
- Verification and Validation (for software and products of ML pipelines)
- Open Source Pipelines and Benchmarks (especially in Operations)
- Smaller is better (produce smaller models, trade off between model accuracy and size)
- Edge computing for ground/space segments.

A summary of the discussion for each point follows below.

Verification and Validation (software).
- The ECSS handbook is widely used in ESA/space. A section thereof for ML models is underway.
- Qualitative validation beyond accuracy or other measures is neede. There are some techniques already in place for this that create some constraints for the model output, also looking at variables not seen by the model.
- Undecidability of the presence of bugs from the achieved accuracy of models. ML corrects for that and helps hide them. Certification procedures should account for these effects.
- We need ontologies of image labels.

Verification and Validation (products).
- New products are derived from ML pipelines (synthetic data). These need to be certified and /or traced back to the originating pipeline/images.
- Detection of fakes might become important to guarantee product value. Might be related to anomaly detection.
- Revise TRL definitions are based on experience from 20 years ago. ML changed all this and they should be reassessed.
- Revise technology trees to account for the change in philosophy coming from ML advances and technological innovations.

Open Source.
- Many projects need to be extended to account for space constraints (Normalization for example needs to happen in a very specific way, often not provided in the OS toolboxes)

- One needs to be careful in general because some data (EO) have a geopolitical value that may drive (wrongly or rightly) political decisions. (example forecast of crop productions). A solution might come on the availability of reference data, or anonymization of data (difficult to do, needs some compromise).
- Complete lack of benchmarks for operation scenarios. A fact. EO is very well ahead w.r.t. other fields in space.
- Culture change needed .... Many agree on the importance of opening up data, but nobody wants to sign documents allowing it.

Edge Computing.

- Importance of having computational constraints accounted for in the development of ML models from early stages.
- Hardware available on board is moving slowly towards higher capabilities but will never close the gap with Earth counterparts. Awareness on application requirements must drive the development of models that fit on foreseeable on board architectures.
- Pruning,distilling, teacher-student models are actively researched areas in ML that should be assessed for space applications.

8.3 Working Group 3

Jurica Levatic (Jožef Stefan Institute – Ljublana, SI)
Sylvain Lobry (Université Paris Cité, FR)
Luke Lucas (LSE Space – Darmstadt, DE)
Jose Martinez-Heras (Solenix – Darmstadt, DE)
Gauthier Picard (ONERA/DTIS, Université de Toulouse, FR)

The working group 3 discussed the following points:

- The loop between planning, execution and monitoring should be closed.
- Anonymization (i.e. adding noise so parameters cannot be reversed-engineered) is something to be investigated.
- In foundation models, standardization and data availability are main issues.
- Existing foundation models could be used for on-board lossless compression or compression with loss (select relevant data to downlink).
- In anomaly detection, the end-user should be involved in the designing of the tool. The use of of AI should be planned from the early stages of mission development.
- In diagnostics, smart filtering, dependency analysis, and XAI are used, but not yet causality. Causal Inference remains a challenge in space operations. It will be very useful for diagnosing anomalies and increasing understanding.

8.4 Working Group 4

Jonathan Bamber (University of Bristol, GB & Technical University of Munich, DE)
Sašo Džeroski (Jožef Stefan Institute – Ljubljana, SI)
Dino Ienco (INRAE – Montpellier, FR)
Ana Kostovska (Jožef Stefan Institute – Ljubljana, SI)
Panče Panov (Jožef Stefan Institute – Ljubljana, SI)

The discussion of the working group 4 can be summarized as follows. Descriptions of the different types of data collected by different EO missions are needed. This would facilitate their use, reuse and combination. This is especially important for applying ML methods to these data and combinations thereof.

- **Ground-truth data is sorely needed.** The same holds for meta-data describing the ground-truth data, which is even more important, because in some cases the data might not be available. Meta-data are essential for finding data relevant to a problem at hand. In addition, it would be very important for transfer learning and learning foundation models. The decision on which foundation model is most relevant for a particular downstream task can be also taken much more competently if we have a description of the data at hand (geographical region, type of urban system/ ecosystem, type of data).
- **Reuse of historical heterogeneous data** (sensors and calibration data) in order to make value of that information for nowcasting/ forecasting is of primary importance and AI can provide methodologies and techniques for such "transfer".
- **Hindcasting is an interesting avenue for further work.** If we have, e.g., both Landsat and LIDAR, for a recent period, we could learn to map forest cover (height and density) from Landsat. We could then get estimates of forest cover for the entire historical period (60 years), where we have Landsat data.
- **Incrementally/Continually updating machine learning models with new EO and calibration data** in order to avoid retraining the system from scratch is also a possible challenge. This is especially important in the context of EO missions that are acquiring systematically new information on the different areas of the Earth surface.
- **Federated learning holds significant potential for EO applications.** The decentralized training methodology would enable us to learn from massive amounts of data without the need of centralizing it. This aspect is particularly advantageous in EO, where data can be voluminous, diverse, and often sensitive w.r.t. privacy and security. Initial work on EO federated learning can be found in `https://arxiv.org/pdf/2311.06141.pdf`. Federated and transfer learning can be used to avoid issues with sharing data. When data cannot be shared, but models can, models can be sequentially (pre)trained and (fine)tuned. The data can stay in place, but models can move around and evolve.
- **Semantic resources for EO need to be developed**, e.g. controlled vocabularies/ ontologies for describing EO-related data, as well as EO-related machine learning tasks, to match to methods.
- **Open publicly funded centre for AI.** There is a high risk that one of the big tech companies creates a monopoly in the field of AI4EO and particularly by the development of very large foundation models. To avoid the risks associated with commercial imperatives that might drive such an approach, publicly funded foundation models are needed.

Participants

- Jonathan Bamber
University of Bristol, GB
- Mitra Baratchi
Leiden University, NL
- Damian Borth
Universität St. Gallen, CH
- Gustau Camps-Valls
University of Valencia, ES
- Nuno Carvalhais
MPI für Biogeochemie – Jena, DE
- Michelangelo Ceci
University of Bari, IT
- Dan Crichton
Jet Propulsion Laboratory – Pasadena, US
- Mihai Datcu
University Politehnica of Bucharest, RO
- Alessandro Donati
Brombachtal, DE
- Sašo Džeroski
Jožef Stefan Institute – Ljubljana, SI
- Simone Fratini
Solenix Engineering GmbH – Darmstadt, DE
- Holger H. Hoos
RWTH Aachen, DE
- Dino Ienco
INRAE – Montpellier, FR
- Dario Izzo
ESA / ESTEC – Noordwijk, NL
- Žiga Kokalj
ZRC SAZU – Ljubljana, SI
- Ana Kostovska
Jožef Stefan Institute – Ljubljana, SI
- Bertrand Le Saux
ESA – Frascati, IT
- Jurica Levatic
Jožef Stefan Institute – Ljubljana, SI
- Sylvain Lobry
Paris Cité University, FR
- George Anthony Long
Legal Parallax – Fountain Hills, US
- Luke Lucas
LSE Space – Darmstadt, DE
- Jose Martinez Heras
Solenix – Darmstadt, DE
- Yazan Mualla
University of Technology of Belfort-Montbéliard, FR
- Jakub Nalepa
Silesian University of Technology – Gliwice, PL
- Evridiki Ntagiou
ESA / ESOC – Darmstadt, DE
- Pance Panov
Jožef Stefan Institute – Ljubljana, SI
- Claudio Persello
University of Twente, NL
- Gauthier Picard
ONERA/DTIS, Université de Toulouse, FR
- Markus Reichstein
MPI für Biogeochemie – Jena, DE
- Jakob Runge
DLR – Jena, DE
& TU Berlin, DE
- Marjan Stoimchev
Jožef Stefan Institute – Ljubljana, SI
- Alexandru Tantar
Luxembourg Inst. of Science & Technology, LU
- Leon van der Torre
University of Luxembourg, LU
- Jan van Rijn
Leiden University, NL
- Joaquin Vanschoren
TU Eindhoven, NL
- Xiaoxiang Zhu
TU München, DE

Report from Dagstuhl Seminar 23462

Defining and Fortifying Against Cognitive Vulnerabilities in Social Engineering

Yomna Abdelrahman[*1], **Florian Alt**[*2], **Tilman Dingler**[*3], **Christopher Hadnagy**[*4], **Abbie Maroño**[*5], **and Verena Distler**[†6]

1 **University of the Bundeswehr – Munich, DE.** `yomna.abdelrahman@unibw.de`
2 **University of the Bundeswehr – Munich, DE.** `florian.alt@unibw.de`
3 **Delft University of Technology, NL.** `t.dingler@tudelft.nl`
4 **Social-Engineer – Orlando, US.** `chris@social-engineer.com`
5 **Social-Engineer – Orlando, US.** `abbie@social-engineer.com`
6 **University of the Bundeswehr – Munich, DE.** `verena.distler@unibw.de`

Abstract

Social engineering has become the main vector for human-centered cyber attacks, resulting from an unparalleled level of professionalization in the cybercrime industry over the past years. Hereby, through manipulation, criminals seek to make victims take actions that compromise security, such as revealing credentials, issuing payments, or disclosing confidential information. Little effective means for protection exist today against such attacks beyond raising awareness through education. At the same time, the proliferation of sensors in our everyday lives – both in personal devices and in our (smart) environments – provides an unprecedented opportunity for developing solutions assessing the cognitive vulnerabilities of users and serves as a basis for novel means of protection.

This report documents the program and the outcomes of the Dagstuhl Seminar 23462 "Defining and Fortifying Against Cognitive Vulnerabilities in Social Engineering". This 3-day seminar brought together experts from academia, industry, and the authorities working on social engineering. During the seminar, participants developed a common understanding of social engineering, identified grand challenges, worked on a research agenda, and identified ideas for collaborations in the form of research projects and joint initiatives.

Seminar November 12–15, 2023 – https://www.dagstuhl.de/23462

2012 ACM Subject Classification Security and privacy–Human and societal aspects of security and privacy

Keywords and phrases Social Engineering, Cognitive Vulnerabilities, Phishing, Vishing

Digital Object Identifier 10.4230/DagRep.13.11.103

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Defining and Fortifying Against Cognitive Vulnerabilities in Social Engineering, *Dagstuhl Reports*, Vol. 13, Issue 11, pp. 103–129
Editors: Yomna Abdelrahman, Florian Alt, Tilman Dingler, Christopher Hadnagy, and Abbie Maroño
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Yomna Abdelrahman (University of the Bundeswehr – Munich, DE)
Florian Alt (University of the Bundeswehr – Munich, DE)
Tilman Dingler (Delft University of Technology – Delft, NL)
Christopher Hadnagy (Social Engineer – Orlando, US)
Abbie Maroño (Social Engineer – Orlando, US)

Social engineering which is defined as "any act that influences a person to take an action that may or may not be in their best interests". In regards to when social engineering is being used by threat actors it is used as psychological manipulation of people into performing actions or disclosing confidential information. Sadly, this form of attack has existed for almost as long as mankind itself. With the advent of AI tools, this form of attack reached a new quality, posing a threat to any online user. Prominent forms of social engineering are phishing attacks and their various subforms (vishing, twishing, QRishing, etc.), physical attacks (dumpster diving, tailgating), and, more recently, deep fakes.

This three-day Dagstuhl Seminar on "Defining and Fortifying Against Cognitive Vulnerabilities in Social Engineering" brought together experts in (user-centered) security, psychology, HCI, computer science, and ethics to identify grand challenges and identify a research roadmap for mitigating social engineering threats. Over the course of the seminar, participants developed an in-depth understanding of the seminar topic. This was achieved by focusing on different aspects of social engineering, discussing how it links to the users' vulnerabilities, namely cognitive vulnerabilities, and how mitigation approaches can be developed.

Day 1 began by introducing the seminar topic, focus, and goals. Afterwards, all participants introduced themselves and their areas of expertise. Each participant contributed and described reading material related to the seminar topic. The material was made accessible to all seminar participants and is attached as a reading list to this report. Following the introductions, day one featured a keynote by Prof. Angela Sasse, entitled "Manipulation, Deception, and Self-Deceit – Broadening Our Perspective of Social Engineering". It highlighted how and why the digital environment makes us so susceptible to social engineering. It took a critical perspective on state-of-the-art approaches to address social engineering. The second talk of day one was given by Chris Hadnagy, who presented important and practical insights into the strategies of modern hackers. Both talks gave a compelling overview of social engineering attacks, an understanding of the most commonly targeted vulnerabilities, and a sense of why it is difficult to mitigate them. Participants then worked in groups to identify grand challenges in social engineering from both researchers' and practitioners' perspectives. Dr. Thomas Kosch and Dr. Yomna Abdelrahman jointly led the last session of day one. It focused on detecting cognitive vulnerabilities and provided an overview of sensing technologies and users' internal states to be inferred, e.g., fatigue, cognitive load, etc. Day one concluded with a group work activity on what we can learn from modern sensors and how to design systems and methods to help mitigate social engineering attacks.

Day two started with a keynote by Mary D'Angelo, addressing the complex topic of understanding and tracing threat actors and social engineers on the dark web. It highlighted the need for collaborative efforts to understand this evolving threat better. Mary D'Angelo and Chris Hadnagy led an open discussion: on the one hand, it focused on the role of practitioners and industry in providing realistic data sets and insights from real-life attacks.

On the other hand, the question of how researchers could use those datasets to (a) better understand attacks and (b) design mitigating techniques was discussed. The second activity on day two was a walk to the ruins, during which participants, led by Claude Kirchner, discussed the ethical aspects of the seminar topic. The afternoon of day two was a group work activity led by Dr. Mohamed Khamis in which participants worked towards addressing the previously identified grand challenges. Breakout groups focused on the different attack phases. Day two ended by transforming the proposed solutions into concrete research projects and agendas.

Day three started with a keynote by Alia Saad, which demonstrated different approaches to addressing human-centered security issues from a technical perspective, using examples from current research. Participants followed up on the proposed research projects in the second session of the day, led by Prof. Florian Alt and Prof. Tilman Dingler. They worked together on refining their ideas and identifying potential collaborations.

This Dagstuhl Seminar provided a platform for interdisciplinary collaboration, fostering a deeper understanding of social engineering and its cognitive vulnerabilities. The identified grand challenges and proposed research projects underscore the importance of collaborative efforts between researchers and practitioners in fortifying against evolving social engineering threats. The insights of this seminar lay the foundation for future research and initiatives in the ongoing battle against malicious psychological manipulation in the digital age.

This seminar had several outcomes. First, it established a community of researchers and practitioners with a common understanding of emerging security threats through social engineering. Second, grand challenges were identified that led to a roadmap for social engineering research, including various research questions addressing theoretical, practical, and methodological aspects. Third, ideas for joint research projects emerged, for several of which an initial consortium was established. Among these projects is the idea of establishing a European Research Center on Awareness, Detection, and Mitigation of Social Engineering, the utilization of a dark web dataset that provides insights into the behaviors of threat actors that lead up to an attack, the utilisation of AI to detect sensitive information in unwanted data disclosures (e.g., via social media shares), and an approach to detecting threats in audio conversations based on voice features and conversation behaviours.

2 Table of Contents

3 Overview of Talks

3.1 Manipulation, Deception, and Self-Deceit – Broadening Our Perspective of Social Engineering

Angela Sasse (Ruhr-Universität Bochum, DE, martina.sasse@rub.de)

The talk examines why the digital environment makes us so susceptible to social engineering – because we have become so used to being manipulated and deceived by others that we don't notice and deceive ourselves that we have choice and control. In fact, ubiquitous tracking of our online activities has created a huge information asymmetry, which Soshanna Zuboff has described as "surveillance capitalism", and humans have adopted routines to respond to prompts to give our time, attention, and money. While we do not regard them as "attackers" in the traditional cybersecurity sense, they utilize very similar cues and exploit habits. To regain control, we need to engage in regular goal setting and planning of our activities and ration our digital engagements along the lines of Cal Newport's "digital minimalism". But we also urgently need reliable trust anchors to enable humans to distinguish friends from foes.

3.2 Physiological Security and Cognitive Vulnerabilities

Thomas Kosch (HU Berlin, DE, thomas.kosch@hu-berlin.de)
Yomna Abdelrahman (University of the Bundeswehr – Munich, DE, yomna.abdelrahman@unibw.de)

Human physiology exerts electric potentials that can be captured by computing devices. Such physiological signals allow the assessment of user states, such as cognitive workload, affect, or stress [2]. While these states can be assessed to allow users to quantify themselves, they are also a gateway for exploiting behaviors in real-time. Social engineering attacks can be tailored depending on the user states, thus increasing the likeliness of social engineering attacks. Furthermore, sensors, such as thermal cameras, are becoming more ubiquitous. Thermal cameras have recently drawn the attention of HCI researchers as a new sensory system enabling novel interactive systems. They are robust to illumination changes, making separating objects from the scene background easy. Far-infrared radiation, however, has another characteristic that distinguishes thermal cameras from their RGB or depth counterparts as it operates in the non-visual spectrum. On the other hand, the visual spectrum, i.e., human visual perception, is limited to only 1 percent of the electromagnetic spectrum. Research has shown that extending visual perception can be beneficial. To investigate the potential of the adoption of thermal imaging, we present the conducted studies to infer users' states, e.g., cognitive load, attention type [5, 6], as well as environmental state, e.g., the presence of recording devices [7], and foot traces [8]. Our findings reflected the potential of thermal imaging to further protect the user by knowing the user's state and nudging them when they are cognitively vulnerable. Yet, our research also explores how thermal imaging might introduce novel attacks, namely thermal attacks[9].

References

1 S. H. Fairclough. 2009. Fundamentals of physiological computing. Interacting with computers, 21(1-2), 133-145.
2 T. Kosch, J. Karolus, J. Zagermann, H. Reiterer, A. Schmidt & P.W. Woźniak. 2023. A survey on measuring cognitive workload in human-computer interaction. ACM Computing Surveys. Association for Computing Machinery, New York, NY, USA
3 R.W. Picard. 2000. Affective computing. MIT Press.
4 N. Sharma & T. Gedeon. 2012. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. Computer methods and programs in biomedicine, 108(3), 1287-1301.
5 Y. Abdelrahman, E. Velloso, T. Dingler, A. Schmidt & F. Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. https://doi.org/10.1145/3130898
6 Y. Abdelrahman, A.A. Khan, J. Newn, E. Velloso, S.A. Safwat, J. Bailey, A. Bulling, F. Vetere & A. Schmidt. 2019. Classifying Attention Types with Thermal Imaging and Eye Tracking. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 3, Article 69 (September 2019). https://doi.org/10.1145/3351227
7 S. Prange, A. Shams, R. Piening, Y. Abdelrahman & F. Alt. 2021. PriView– Exploring Visualisations to Support Users' Privacy Awareness. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445067
8 A. Saad, K. Izadi, A. Khan, P. Knierim, S. Schneegass, F. Alt & Y. Abdelrahman. 2023. Hot-Foot: Foot-Based User Identification Using Thermal Imaging. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3580924
9 Y. Abdelrahman, M. Khamis, S. Schneegass & F. Alt. 2017. Stay Cool! Understanding Thermal Attacks on Mobile-based User Authentication. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). Association for Computing Machinery, New York, NY, USA, 3751–3763. https://doi.org/10.1145/3025453.3025461

3.3 Threat Actors and Threat Intelligence on the Dark Web

Mary D'Angelo (Searchlight Cyber – Washington, DC, US)

This speech addresses the complex topic of threat actors and social engineering on the dark web, highlighting the need for collaborative efforts to understand this evolving threat. The speaker begins by providing a historical overview of the development of the dark web, from the creation of ARPANET in 1969 to the advent of TOR and Bitcoin, which have facilitated a surge in dark web activities over the last 15 years. Current trends in the dark web, including the rise of malicious social engineering practices, are discussed, with examples such as services for phishing, vishing attacks, and educational resources for threat actors. The MGM hack by Scattered Spider serves as a case study to illustrate the sophisticated nature of these attacks. The speaker emphasizes the urgency of understanding the mechanisms of threat actor communication and transaction on the dark web, the organization of these actors, and their growing capabilities, as evidenced by a significant increase in vishing attacks. The speech concludes with a call for collaborative research between practitioners and the academic community to develop effective defenses against these evolving cyber threats.

3.4 Biometrics Against Social Engineering

Alia Saad (University of Duisburg-Essen, DE, alia.saad@uni-due.de)

This talk demonstrated several approaches to mitigating human-centered attacks based on current research examples. The talk was meant to inspire discussion among participants as to how the challenges identified during the seminar can be approached in joint research projects. The first part demonstrated how situations in which users are exposed to a human-centered attack can be studied in detail, using shoulder-surfing as an example [1]. Furthermore, an example was shown of how a user interface can be built that points out risk in-situ [2]. The second part demonstrated how the need for user interaction can be minimized by creating, implementing, and evaluating technical approaches seamlessly running in the background, using behavioral biometrics as an example. The talk demonstrated how mechanisms based on different behaviors can be built, in particular, gait [6] and hand-based interaction [3]. Furthermore, the talk also demonstrates a system for use in everyday life [4].

References

1 A. Saad, J. Liebers, U. Gruenefeld, F. Alt & S. Schneegass. 2021. *Understanding Bystanders' tendency to shoulder surf smartphones using 360-degree videos in virtual reality.* In Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction (MobileHCI '23). Association for Computing Machinery, New York, NY, USA.
2 A. Saad, M. Chukwu & S. Schneegass. 2018. Communicating shoulder surfing attacks to users. In Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia (MUM '18). Association for Computing Machinery, New York, NY, USA.
3 A. Saad, M. Pascher, K. Kassem, R. Heger, J. Liebers, S. Schneegass & U. Gruenefeld. 2023. Hand-in-Hand: Investigating Mechanical Tracking for User Identification in Cobot Interaction. In Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia (MUM '23). Association for Computing Machinery, New York, NY, USA.
4 A. Saad, K. Izadi, A.A. Khan, P. Knierim, S. Schneegass, F. Alt & Y. Abdelrahman. 2023. HotFoot: Foot-Based User Identification Using Thermal Imaging. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 262, 1–13. https://doi.org/10.1145/3544548.3580924
5 Mihai Bâce, Alia Saad, Mohamed Khamis, Stefan Schneegass, and Andreas Bulling. *PrivacyScout: Assessing Vulnerability to Shoulder Surfing on Mobile Devices.* Proceedings on Privacy Enhancing Technologies 3 (2022): 650-669.
6 A. Saad, N. Wittig, U. Gruenefeld & S. Schneegass. 2022. A Systematic Analysis of External Factors Affecting Gait Identification. In the IEEE International Joint Conference on Biometrics (IJCB), pp. 1-9. IEEE.

4 Working Groups

Our seminar brought together participants from academia, industry, and the authorities. We began our seminar with level-setting. Short foundational talks from practitioners and researchers aimed to foster a common level of understanding of the differing perspectives of the various communities as well as a joint language. Based on this, the seminar focused on interactive formats (break-out groups, open discussions) intending to identify grand challenges, a research roadmap, and opportunities for collaboration.

4.1 Manipulation Mastery – The Strategies of Modern Hackers

Chris Hadnagy (Social Engineer – Orlando, FL, USA, chris@social-engineer.com)
Abbie Marono (Social Engineer – Orlando, FL, USA, abbie@social-engineer.com)

Objective Due to the very different backgrounds of seminar participants (scientists, practitioners, military), this session aimed to create a common understanding of social engineering and identify state-of-the-art strategies of hackers.

Methodology This session followed a two-step approach to develop a common understanding. Firstly, Chris Hadnagy introduced the fundamentals of social engineering. Afterward, participants were split into groups, discussing a set of questions to create a joint understanding of social engineering from different perspectives.

Step 1: Fundamentals

The talk by Chris Hadnagy aims to define the fundamentals and main vectors used by malicious social engineers in attacking their targets. We define how attackers use phishing, vishing, SMiShing, and impersonation in their attacks. By defining each of these and discussing the advancement in the technology used, we can better understand the minds of the attackers in choosing which method to use. Following this discussion, we went into depth about the stages of social engineering engagement from a practitioner standpoint. The goal was to understand the methodology used by professional social engineers.

4.1.1 Terminology: Social Engineering

Social engineering involves manipulating individuals to give away confidential information or perform actions that compromise security. Techniques include phishing, vishing, and impersonation, where hackers exploit human psychology and vulnerabilities of different kinds.

4.1.2 Attack Vectors

State-of-the-art attack vectors encompass a range of techniques malicious social engineers utilize to exploit vulnerabilities and gain unauthorized access to sensitive information or systems. Prominent attack vectors include:

Phishing Phishing involves fraudulent attempts to obtain sensitive information, such as usernames, passwords, and credit card details, by disguising as a trustworthy entity in electronic communication. These attacks commonly occur via email, where users are persuaded to click malicious links or provide confidential information.

Vishing Vishing, or voice phishing, is a form of social engineering where attackers use phone calls to deceive individuals into disclosing personal or financial information. The attackers may impersonate legitimate organizations or individuals to manipulate victims into revealing sensitive data or performing certain actions.

Smishing Smishing, or SMS phishing, exploits text messaging systems to trick users into revealing personal information or installing malware on their devices. Attackers send deceptive text messages containing malicious links or requesting sensitive information, exploiting the trust associated with SMS communications.

Impersonation Impersonation attacks involve masquerading as a trusted entity, such as an authority, figure, or reputable organization, to deceive individuals into disclosing confidential information, transferring funds, or performing actions that compromise security. Attackers often use social engineering tactics to gain the trust of their targets before exploiting it for malicious purposes.

It is worth mentioning that these attack vectors continue to evolve as attackers adapt their strategies to bypass security measures and exploit new human vulnerabilities.

Step 2: Breakout Groups

Following the introduction and plenum discussion to obtain a common understanding, participants split into two subgroups to discuss the following questions:

4.1.3 Guiding Questions

- How do modern hackers target their victims?
- How are you educating yourself/ your team on security awareness?
- What are the limitations of this type of education?
- What problems have you encountered trying to manage threats from Social Engineering?
- What kind of collaborations/technologies would help increase your security?

4.1.4 Outcomes

The groups collated a range of techniques of modern attacks, including:

Deepfake Technology Attackers may use deepfake technology to create convincing fake audio or video recordings for social engineering attacks, impersonating trusted individuals or manipulating content.

Machine Learning and Artificial Intelligence As technologies evolve, attackers may leverage machine learning and artificial intelligence to enhance their attacks. This includes creating more sophisticated malware, evading detection, or automating certain aspects of the attack process, e.g., generating phishing emails.

Augmented Reality (AR) and Virtual Reality (VR) Threats Attackers may exploit AR and VR technologies for social engineering attacks, creating immersive scenarios to deceive victims or launching attacks within virtual environments.

When the groups discussed the limitations of training or educating users and the problems of managing social engineering attacks, they came up with several reasons why this can be challenging, including the following:

It is difficult to convince people that damage is possible and real for several reasons:

- **Lack of personal experience:** Users may not have personally encountered an incident, leading to a perception that such events are rare or unlikely to affect them. Without direct experience, it can be challenging to grasp the potential consequences.
- **Trust in technology:** We often trust and rely on technology in our daily lives, which might make us feel that systems are secure and that incidents like data breaches or cyberattacks won't happen to us.
- **Threats are invisible:** Unlike physical attacks resulting in visible damage, the effects of social engineering may be hidden. Data breaches, e.g., might not immediately manifest as tangible harm, making it difficult for individuals to recognize the severity of the situation.
- **Reaction times are generally too long:** Victims may feel overwhelmed when confronted with the potential damage caused by falling for a social engineering attack. This discomfort can lead to denial or avoidance of reporting the issue.

Training does not work because humans forget / cannot memorize everything: One-time or infrequent training sessions may not be sufficient to create lasting awareness.

Current mitigation strategies are difficult to scale: Several factors contribute to the difficulty of scaling social engineering mitigation strategies. The discussed challenges covered:

- **Limited resources:** Many organizations have limited resources, both in terms of time and budget, to devote to extensive training programs.
- **Constantly evolving attacks:** Social engineering strategies continuously evolve, and attackers regularly develop new techniques. Staying ahead of these evolving threats and updating training content accordingly is resource-intensive and time-consuming.
- **Measuring effectiveness:** Determining the effectiveness of social engineering training programs is challenging.
- **Privacy concerns:** Balancing the need for effective social engineering training with respect for users' privacy can be a delicate task. Some individuals may hesitate to participate in training that they perceive as invasive.
- **Existence and scalability of Technical Solutions:** Implementing technical solutions to detect and prevent social engineering attacks can be complex and do not exist yet.
- **Lack of ecological validity:** The groups discussed potential collaboration between practitioners and researchers to address the above-mentioned challenges better. The reported concern was the struggle of researchers to move research out of the lab; current approaches are often scenario-oriented (but lack the ecologic validity of real-world settings).

4.2 Grand Challenges in Social Engineering

Matteo Große-Kampmann (Aware7 – Gelsenkirchen, DE, matteo@aware7.de)
Angela Sasse (Ruhr-Universität Bochum, DE, martina.sasse@rub.de)

Objective This session aimed to identify grand challenges in social engineering from both a practical and academic perspective.

Methodology To structure this working group, Chris Hadnagy first introduced the different phases leading up to a successful social engineering attack. Those steps then served as a scaffold for breakout groups in which grand challenges for each phase were identified.

Phases of a Successful Social Engineering Attack

Phase 1–OSINT Open Source Intelligence (OSINT) refers to collecting and analyzing information from publicly available sources. It involves gathering data from various sources such as social media, news articles, online forums, public databases, and websites. OSINT provides social engineers with valuable information about their targets, which can be used to craft convincing narratives and exploit vulnerabilities. By utilizing OSINT, social engineers can gather personal details, interests, affiliations, and even behavioral patterns of their targets. This information enables them to tailor their approaches to appear more trustworthy and increase the chances of successful exploitation.

Phase 2–Target Selection Social engineers target individuals who can access valuable information or sensitive systems. This could be employees of a company, individuals in positions of authority, or those with access to financial information. By targeting those with valuable data, social engineers increase the likelihood of a successful attack or fraud.

Phase 3–Attack Plan Social engineers craft a highly personalized attack plan using the information gathered during the prior phases. They adopt different personas, using tactics such as impersonation, pretexting, or creating fake online profiles to establish credibility. They exploit emotions and trust by pretending to be someone the target knows or trusts.

Phase 4–Attack Launch Social engineers conduct their attacks by manipulating and exploiting human psychology and trust. They use various tactics to manipulate individuals into divulging confidential information or performing actions that could compromise security. They often employ techniques like impersonation, pretexting, and phishing to trick people into believing they are someone they are not or representing a trustworthy entity. Social engineers can access sensitive information and financial data by exploiting human emotions, curiosity, and ignorance or by gaining unauthorized entry into systems. These attacks can occur through various mediums, such as phone calls, emails, social media, or even in-person interactions, to deceive individuals and bypass security systems.

Phase 5–Evaluation Throughout the process, social engineers document their actions, record findings, and assess the impact of any successful exploits.

Phase 6–Reporting Finally, they provide a detailed report to the organization, outlining the vulnerabilities discovered and recommending remediation measures.

Pratical Challenges and Research Challenges

Phase 1–OSINT

Participants discussed why educating users to protect themselves from the initial phase of social engineering is complex. The participants identified the following challenges concerning Open Source Intelligence:

Creating Awareness of Own Vulnerabilities: A key challenge is making users aware of their vulnerabilities. Users should be aware of what information is publicly shared and could be used by attackers (and which information is not publicly stored but could be accessible by attackers if the platform is breached). Moreover, many users struggle to think that they are targets in the first place.

Inference of Available Data: A challenge is keeping an overview of available information about oneself when AI models draw conclusions based on metadata (e.g., relationships).

Social Media Exploitation: Social media platforms often contain tons of personal information. Attackers can exploit this information to create phishing messages, impersonate individuals, or conduct other forms of social engineering.

Dynamic and Evolving Nature of Platforms: Online platforms and available information constantly change and evolve. Keeping track of these changes and assessing the reliability of information can be challenging, for instance, how social media platforms change their privacy settings and auto-send friend requests.

Lack of Users' Awareness of the Influence of Their Internal State: Users' internal states and vulnerabilities play a significant role in their susceptibility to social engineering attacks. As social engineering relies on manipulating individuals' emotions, behaviors, and cognitive processes to deceive them into revealing sensitive information, taking specific actions, or compromising security. Yet, users are not aware of such an influence.

Based on the group discussion and the pointed-out challenges, participants discussed potential solutions to help protect the user and increase their awareness about OSINT. Attackers usually use information aggregation during the OSINT phase, where they gather information from multiple sources to create a comprehensive profile of a target. This profile can be used to craft more convincing and targeted social engineering attacks. Accordingly, participants envisioned a *Cross-Platforms Search Notification System*, where users would be alerted if someone searches them on different platforms, e.g., work/personal website, LinkedIn, Instagram, Facebook, and warn the user if access rates are unusually high. However, this entails technical and privacy challenges, e.g., logging search activities across platforms.

Phase 2–Target Selection

Identifying potential targets and implementing strategies to mitigate risks can be complex for both defenders and users. Participants openly discussed why this phase might be hard to mitigate. One dominant reason was that many users struggle to think they were targets in the first place. Hence, they exhibit neglectful behavior when dealing with both their own data and institutionally accessible information.

Phase 3–Attack Plan

The central premise of an attack plan is coming up with a pretext for the attack. During the planning phase, the attacker uses the information gathered from OSINT to devise a strategy. This involves selecting the most appropriate attack vector—whether it be phishing, pretexting, baiting, or another method—based on the target's vulnerabilities and the attacker's objectives. The attacker also crafts the message or scenario they will use to deceive the target, ensuring it is convincing enough to elicit the desired response. This phase requires careful consideration of the psychological and emotional triggers that will be most effective on the target and planning for any contingencies or responses the target might have.

Attack planning also involves the creation of backstories, fake identities, or any necessary props (like counterfeit badges or websites) that will make the attacker's approach more credible. This is where the creativity and insight of the attacker into human psychology are paramount. The success of this phase hinges on how well the attacker can anticipate the target's reactions and prepare for them, ensuring that the attack will not only reach the target but also resonate with them, prompting the desired action or information disclosure.

One of the most effective ways to counteract the planning phase is to limit already access to information that could be gathered during OSINT. Corporations should regularly check what type of information is publicly accessible, including about their employees. Awareness campaigns about public profiles can hone employees' sensitivity to sharing certain information about themselves or their employers.

Phase 4–Attack Launch

In the attack launching phase, the attacker puts their plan into action and makes direct contact with the target. This could be through email, phone calls, social media, or in-person interactions. The attacker employs the crafted scenario to manipulate the target into performing specific actions, such as divulging sensitive information, granting access to secure systems, or even transferring funds. The success heavily relies on the attacker's ability to adapt to the conversation flow and maintain the deception convincingly.

During this phase, the attacker must remain vigilant and adaptable, as unforeseen variables or responses from the target may require on-the-fly adjustments to the plan. The psychological manipulation skills of the attacker are crucial here, as they must build trust or authority with the target quickly. The ability to read cues from the target and adjust the approach accordingly can make or break the attack's success.

Countermeasures can be taken individually through general awareness training or collectively through peer protection. The latter entails establishing a reporting culture in which employees inform and warn each other about new schemes they encounter. Employers should establish a single point of contact where incidents can be easily reported and that is responsible for disseminating newly identified threats. Another approach that has merit on both individual and collective levels is the introduction of *friction*. Artificially delaying certain actions or procedures, e.g., can create time windows in which reason can kick in, or the attack can be delayed to a point where the risk of exposure becomes too great to continue the attack. Urgency should almost always be a warning sign for an incoming attack.

Phase 5–Evaluation

Evaluating social engineering attacks, particularly those conducted as part of penetration testing, poses several challenges.

Ecologic Validity and Generalizability While pen testers try to act realistically, their actions are still limited by legal and ethical considerations. Also, their customers might exclude certain actions (e.g., accessing sensitive and personal information about employees). Those conditions inevitably influence the ecologic validity of the findings and pose the question of to which degree the findings generalize to settings with real attackers.

Metrics Many forms of penetration testing are still strongly limited in terms of the used metrics. For example, click rates are among the most popular metrics for phishing awareness campaigns. At the same time, these allow only very little to be learned and are questionable, as they depend on factors beyond pen testers' control (and assessment). There is a need to rethink metrics currently in use fundamentally.

Individualization Measures against social engineering are generally costly from a corporate perspective, as a result of which easy-to-implement solutions are favored (e.g., making users attend talks on awareness once a year). The challenge with such measures is that they might annoy users (as content might be repetitive). Also, employees might have a different level of knowledge and understanding, as a result of which some might struggle with terminology already while others might be bored. A major challenge is the individualization of measures, where users' skills, prior knowledge, and tasks of their everyday job are considered.

Case Study vs. Large Scale Due to the required effort and cost, campaigns and research projects often focus only on specific cases rather than large-scale approaches. While (small-scale) case studies might be well suited to identify causes and interesting aspects, more large-scale studies are required to assess effects.

Priming For ethical reasons, users or employees might be primed; that is, they are being told upon being employed or signing up for a study that they will / might be subject to a security assessment. This inevitably changes behavior. Research is needed to understand the implications of priming and on approaches that minimize any such priming effects.

Independent Auditing Pentesting / auditing is often conducted due to certification or due to being required by insurance companies. As a result of this, companies subsequently hire auditors. The challenge here is that pen tests and audits, in this case, are not independent. It is still an open question about how such independence can be achieved.

Completeness While defenders must protect against any vulnerability, attackers only need to find one weakness. Pentesting usually cannot consider any aspect/attack surface.

Pentester Empathy A particularly interesting aspect is their actions' implications on pen testers (similar to the Milgram experiment). This is an unexplored area of research.

Phase 6–Reporting

The group identified several challenges regarding the reporting and, in particular, the way in which recommendations are / should be made.

Turning lessons learned into positive change While many pen tests are designed to demonstrate issues/holes that allow attackers to be successful, it is often much less clear how this knowledge can be turned into positive change.

Targeting Opportune Moments Closely related to the abovementioned aspect, change must be carefully targeted to opportune moments, i.e., moments in which users are (more) receptive to change and appreciate it. It is well known that in situations of change (e.g., moving to a new house/office, getting a new smartphone/laptop), people are more willing to change habits, but this is hardly explored from a cybersecurity perspective.

Check-the-Box vs. Organizational Change Penetration tests/auditing is often seen today as a necessary requirement rather than an opportunity for real (organizational) change. It remains an open challenge to move from just getting things done to a culture in which true organizational change is anticipated.

Misconceptions about being a target Many users struggle to understand/accept that they are a target. Reports can surface convincing cases (e.g., kindergartens being targets of cyber attacks).

Response Costs / Prioritization Cyber attacks are usually possible through many different approaches, and addressing all of them is costly. There is a need to understand better how countermeasures can be prioritized so as to maximize their impact.

Change Management / Leadership Who should drive change is often unclear. Whereas employees often consider employers responsible for cybersecurity, employers want employees to change their habits. A challenge is how to establish a social contract.

Expressing IT Security as a Business Risk In particular, companies and individuals struggle to accept IT security as a business risk until they become victims. Research is needed as to how the consequences of successful cyber attacks can be better conveyed.

4.3 Towards Solutions: Cognitive Vulnerabilities

Thomas Kosch (HU Berlin, DE, thomas.kosch@hu-.berlin.de)
Yomna Abdelrahman (University of the Bundeswehr – Munich, DE, yomna.abdelrahman@unibw.de)

Objective The objectives of this session were to (1) understand how modern sensing technology can be used to assess user states affected by social engineering, to (2) think about how knowledge of those states can be used to design counter measured and (3) how attackers can exploit this knowledge.

Methodology Thomas Kosch and Yomna Abdelrahman first introduced the capabilities of modern sensors and machine learning (see talk). Afterward, participants were divided into groups and asked to think about ways of using knowledge obtainable from sensors to defend from social engineering and also which novel attack surfaces this creates.

As previously discussed, users' internal states significantly influence their susceptibility to social engineering attacks. Social engineering relies on manipulating individuals' emotions, behaviors, and cognitive processes to deceive them into revealing sensitive information, taking specific actions, or compromising security. In this session, we leverage design fiction methods to ideate the role of physiological sensors in social engineering from both the attackers' and defenders' perspectives.

What can we learn from modern sensors?

Thomas Kosch and Yomna Abdelrahman gave concrete examples of utilizing novel ubiquitous sensors like eye tracking and thermal cameras to detect and leverage cognitive vulnerabilities during social engineering. However, opportunities are not limited to these examples but rather to steer the mindset of the participants towards using modern sensors in different contexts.

Eye Tracking: Eye-tracking data in social engineering refers to collecting and analyzing information about a person's eye movements and gaze patterns. While traditional uses of eye-tracking are often associated with research in psychology and HCI usability studies, the application of eye-tracking data in the realm of social engineering introduces additional considerations. Here are some examples:

- **Understanding Attention:** Eye tracking data can provide insights into where a person directs their attention. In a social engineering context, understanding what elements or cues attract a person's gaze can be valuable for attackers. Attackers may use this data to refine deceptive techniques.
- **Phishing and Visual Deception:** Attackers may leverage eye-tracking data to optimize the design of phishing emails. By understanding where users focus their attention, attackers can create more convincing and visually deceptive elements to increase the likelihood of successful social engineering attacks.
- **Defensive Techniques:** While attackers could utilize eye-tracking, they also hold merts for defenders. Researchers and defenders can use eye-tracking data to understand how users visually engage with social engineering attacks and security warnings. This information can inform the design of more effective alerts and communication strategies to raise awareness about potential social engineering threats.

Thermal Cameras: Facial temperature can be influenced by emotional states, and temperature changes may reflect variations in emotions and internal states, e.g., cognitive load, stress, anger, etc. These changes could be monitored seamlessly and non-invasively using thermal cameras. While researchers utilized thermal cameras to build cognitive-aware systems in various contexts, it is unexplored in the context of social engineering, yet the potential it holds, for instance:

- **Cognitive Load Detection:** Recent work showed the potential of using thermal cameras to capture facial temperature to quantify cognitive load from low, medium, high, and very high. Research reflected the potential of thermal imaging to further protect the user by knowing the user's state and nudging them not to perform any security critical tasks when they are cognitively vulnerable.
- **Emotions Detection:** Attackers rely on the victims' emotions to conduct social engineering attacks. Namely, they aim to simulate emotions, e.g., fear, guilt, and stress, to make users reveal sensitive information or perform certain actions. Research revealed correlations between changes in facial temperature and these emotions. Emotional-aware systems built using thermal cameras could detect these emotions and either used by the defenders to build protective measures when such emotions are detected or by the attacker to know vulnerable moments.

Following the talk, we had an open discussion on how modern sensors could be deployed in the context of social engineering to serve both attackers and defenders. To this end, participants split into two subgroups to discuss the following questions:

4.3.1 Guiding Questions

- Imagine you are an IT security designer. How would you use psychological real-time data to improve user security?
- Imagine you are a hacker who has access to psychological real-time data. How would you utilize the data for a social engineering attack?
- How can users be made aware of their individual cognitive vulnerabilities?

How can we use the knowledge of user states to build better protection mechanisms?

The groups came up with the following ideas for novel protection mechanisms.

Mechanism 1 During face-to-face situations, use a video-based assessment of physiological data to give insights into the current level of fatigue, stress/arousal, and identify health status. Based on the signals, the system would recognize if the user is sleepy/exhausted, provide recommendations, and flag potential threats.

Mechanism 2 During reading emails, use data such as heart rate, EDA, pupil dilation, blink rate (fatigue), reading speed, eye tracking for speed of reading emails, and nonverbal body posture to determine when the user is vulnerable, and the email client color changes to indicate cognitive vulnerability.

Mechanism 3 One group proposed using physiological data not to detect victims' vulnerabilities but the attacker's intent. Once an attacker is detected, it is flagged.

How can attackers exploit knowledge of user states?

The groups came up with the following ideas for novel possible attacks.

Attack 1 Attackers could exploit the knowledge about users' states to tailor the attack and abuse the user during their vulnerable state.

Attack 2 Having access to heart rate, EDA, pupil dilation, blink rate (fatigue), and reading speed data, attackers could use the data to find a point when the target is most stressed and fatigued, being overly busy. During that time, a phishing email is sent from someone in authority over the target requesting immediate action.

Interestingly, the participants discussed how they can benefit from the potential of using physiological real-time data without the risk of falling into the attackers' hands. For instance, participants proposed randomly introducing noise to the data or using differential privacy. Adding enough noise makes the attack infeasible but still allows data to be used legitimately.

4.4 Social Engineering for Good (A Walk to the Castle Ruins)

Tilman Dingler (TU Delft, NL, t.dingler@tudelft.nl)
Claude Kirchner (INRIA Institut National de Recherche en Informatique et en Automatique – Rocquencourt, FR, claude.kirchner@inria.fr)

So far, social engineering has been mainly discussed in association with deceptive practices aimed at exploiting human psychology for malicious ends. The goal of this session, however, was also to consider the use of its principles and mechanics for positive individual and social outcomes, *i.e.* for good. Social engineering (SE) can be defined as "any engineered act that influences a person to take an action that may or may not be in their best interest". This allows us to insist on the specially designed intention (engineered) as well as to emphasize that it could be for malicious but also positive "for good" reasons. In this context, we should consider the involved ethics, understood as the thinking process about human conduct and the values on which they are based. Indeed, either for good or bad, social engineering may not respect human autonomy, transparency, or explainability, and of course, the non-maleficence principle will be strongly questioned.

During a joint walk up to the old castle ruins, seminar participants were thus presented with two leading questions and invited to discuss in present company and eventually report back. The two leading questions posed were:

1. How would you use insights, techniques, and methods of Social Engineering to do good?
2. What are ethical boundaries and obligations when "manipulating" people for good?

Examples discussed by participants included personal health and environmental conservation. In the realm of public health, social engineering and, specifically, nudging can play pivotal roles. For instance, designing environments that subtly encourage physical activity, such as strategically placed stairs over escalators, can significantly impact public health outcomes. Similarly, nudges in cafeterias or grocery stores, like placing fruits and vegetables at eye level, can make healthy food choices more appealing and accessible. These interventions use our natural tendencies and decision-making shortcuts for positive ends, making the healthier choice, the easier or more attractive option.

On another note, environmental sustainability can benefit from social engineering techniques aimed at encouraging eco-friendly behaviors. For example, utility companies have successfully used social norms to influence behavior by showing customers how their energy consumption compares to their neighbors, nudging them to reduce energy use. Similarly, simple prompts or reminders to recycle or making recycling bins more visible and accessible can significantly increase recycling rates. These strategies rely on our innate desire to conform to social norms and our responsiveness to environmental cues, guiding us toward more environmentally sustainable actions.

A commonly mentioned critique of nudges was the user's agency, which might potentially be violated. Even nudges "for good" are construed by a choice architect, *i.e.* an individual or group of people who deem one choice *better* than another. Conflicting moral and value systems can, therefore, give precedence to choices that go against what the individual might have selected in a more conscious choice scenario. In the end, any technique deemed as social engineering entails some kind of manipulation. The question of which manipulation is deemed "good" or "bad" needs to be discussed in light of differing moral and ethical frameworks. People's agency should, at best, be preserved, while transparency should always be provided about how certain choices are presented.

4.5 Identifying Research Areas and Research Questions

Mohamed Khamis (University of Glasgow, UK, me@mkhamis.com)

Objective This session aimed to identify specific research questions that could be addressed through joint research projects and initiatives.

Methodology Based on the grand challenges and discussions from the first day of the seminar, Mohamed Khamis synthesized different areas of research participants could vote on. Afterward, breakout groups were built where people identified specific research questions based on their interests.

We compiled a list of *main research areas* based on the outcomes of previous sessions. Participants then voted on which areas they would like to explore. The main areas were:

1. Social engineering vulnerability: self-assessment and misconceptions
2. Organizational changes to defend against social engineering
3. Frictions and warnings: How? When? What?
4. Evaluation of solutions: threats to validity
5. Evaluation metrics
6. Datasets

All areas that received four or more votes proceeded to the next stage, in which the participants chose the area they were interested in exploring further to produce research questions that can be addressed by (a) a Ph.D. thesis, (b) a research grant, or (c) a dedicated research center.

Research Areas

The participants voted for the following research areas:

Evaluation of Solutions The breakout group discussed methodological challenges of evaluating solutions against social engineering attacks.

Finding and Supporting Routines Against Attacks This breakout group discussed research questions related to the development of routines, aiming at minimizing risks from social engineering attacks.

Datasets This breakout group aimed to identify research questions that could be answered by having access to different practitioners' datasets.

Research Questions

4.5.1 Evaluation of Solutions

This breakout group focused on the methodological challenges of evaluating solutions against social engineering attacks, namely phishing and vishing. The group discussed and identified the following research questions.

How can we collect data on successful and unsuccessful attacks? One of the major challenges is the limited access to realistic, ecologically valid data. Access to such data is restricted due to legal and privacy constraints. Additionally, users may exhibit reporting bias, i.e., they may be hesitant to report social engineering attacks, whether successful or unsuccessful. Accordingly, the data collected may be skewed, and the true impact of social engineering attacks may be underestimated. While researchers try to overcome this challenge by relying on simulated or controlled environments, this usually does not fully capture the actual attacks.

How can we identify and support protection strategies among real users? Another interesting research question is how to crowd-source protection strategies from users' common practices. While reporting bias is one challenge, another entailed challenge is how to develop protection strategies that are adaptable and customizable to the diverse needs and characteristics of different user groups. Furthermore, establishing effective feedback loops for users to report social engineering attacks or provide input on protection measures and strategies might be methodologically challenging.

How can we run evaluations in different attack phases? Participants categorized the solution space into three phases: pre-, during-, and post-attack. This categorization introduces a set of research questions: How can attacks be reliably detected? What are effective intervention designs for the different phases? What are the appropriate research methods for each phase? When are solutions most effective?

4.5.2 Finding and Supporting Routines Against Attacks

While much knowledge and tools exist that can help protect users from social engineering attacks, a prevalent challenge is establishing secure routines. An example is using a password manager whenever logging into a website, as, in this way, links to fake phishing websites would be easily identified. The group identified the following research questions.

What is the role of routines? As a first step, researchers could explore the role routines could play in users' everyday lives.

Which routines work across contexts? A challenge is that routines (think about using VPNs) might work in one context, for example, during a business trip – but not in other contexts, such as being on vacation. The development of routines would benefit from an in-depth understanding of which routines work across contexts and which do not.

How to design cues/reminders of risks/security behavior? How to communicate them? To support the development and habituation of routines, an interesting question is how users can be reminded of them, particularly as they are about to behave insecurely/riskyly.

How can the community support routines? Another interesting question is the role of the community, particularly the question of how the fact that a community agreed on certain routines would affect the individual.

What are easy routines? Some routines are easier to habituate than others. Identifying easy routines would be valuable information to support self-efficacy. Routines that are easy for one user might not be easy for another user.

How can routines be supported through AI? Participants of the breakout groups also discussed the question of whether routines could be supported through AI, for example, models that predict opportune moments.

How can developing a security mindset be supported? An interesting question is how the gradual establishment of routines might ultimately lead to a "security mindset" among users and whether this makes adopting routines for other security contexts more likely.

4.5.3 Data Sets

The breakout group on data sets identified questions that could be answered as researchers have access to (historical and real-time) information on threat actor communication, traceable transactions on the dark web, knowledge of the organizations of these actors, and observable actions (e.g., increasing network traffic towards potential victims).

What are observable attacker movements? First, a comprehensive understanding of attacker actions, their characteristics, and how they could be tracked would be interesting.

How can we associate movements with attackers? A current challenge is linking observable movements with particular threat actors, as these are often difficult to identify (due to using TOR, VPNs, etc.). At the same time, close temporal or spatial proximity might hint at movements being associated with particular threat actors, allowing a more comprehensive picture to be drawn.

How can predictive models for attackers from "movement sequences" be built? Researchers were particularly excited about the ability to not only understand what common sequences of movements are but to, based on this knowledge, predict the next steps of attackers. This would give potential victims time to prepare their defense and expect attacks.

What are opportune moments to intervene? From a practical point of view, an interesting question is when to intervene; that is, when to approach and warn potential victims. A predictive model might hint at a particular time window in which an attack is likely.

What should interventions look like? As an attack is likely, an interesting question is how to intervene. Should potential victims be sensitized? Or should they be trained through fake campaigns to (at least temporarily) improve their detection skills?

When should a pen test be run? Along the same lines, an interesting question could be when to launch a penetration test so as to test defenses.

How can we identify opportunities for attackers (victims are unaware of)? Participants found the idea of learning more about the attackers and what opportune moments they exploit. In that way, a better understanding of threats can be obtained that might help victims develop better routines that minimize opportunities for attackers.

4.6 Towards Collaborations in Social Engineering Research

Florian Alt (University of the Bundeswehr – Munich, DE, florian.alt@unibw.de)

Objective This session aimed to identify specific topics and areas of collaboration.

Methodology For this session, a poster was created for each research question participants identified in the previous session. Then, participants were asked to indicate their interest in each research question by writing their names on the poster. Afterwards, in several rounds, participants met at the poster to discuss the following questions: (1) Who would fund this research? What would be the scope of a project? (3) How would you pitch the topic (i.e., write an abstract)?

The following list describes the different research projects and initiatives.

A European Research Center for Awareness, Detection, and Mitigation of Social Engineering

Funding NL: NCSC-NL (National Cyber Security Center), NCTV (National Coordinator For Counterterrorism and Security, Ministry of Justice and Security), Cyberveilig Nederland, HighTechCrime Police/Europol (i.e., Law Enforcement), AIVD (General Intelligence and Security Services, Ministry of Interior)
FR: ANSSI (National Cyber Security Center), Viginum (Service of vigilance and protection against foreign digital influence), Inria (National Research Institute on Informatics and applied mathematics, CNRS (National Center for Scientific Research).
DE: BSI (Bundesamt für Sicherheit in der Informationstechnik), BMI (Bundesministerium des Inneren), Cyberagentur; BKA (Bundeskriminalamt)

Pitch Social engineering attacks have emerged as a predominant threat, exploiting human psychology to manipulate individuals into revealing confidential information, compromising their financial security, and influencing their decision-making, including voting behavior. These tactics bypass traditional cybersecurity measures by directly targeting the most vulnerable link in the security chain: the human. Recognizing the gravity and complexity of this issue, the proposed *enter title here* aims to serve as a pioneering institution dedicated to combating these threats across Europe.
The center will be a collaborative hub, uniting industry practitioners and academic experts to develop comprehensive strategies against social engineering attacks. Its primary objectives will include raising awareness about the nature and tactics of these attacks, enhancing the ability to detect and identify such threats promptly, and devising effective mitigation strategies to reduce their impact. By fostering a multidisciplinary approach, the center will address current challenges and anticipate emerging trends in social engineering, such as the threat of generative AI, ensuring a proactive stance against these evolving threats.
This research center will formulate goals to fortify individual privacy, financial integrity, and democratic processes against the influence of social engineering. The establishment of this research center represents a significant step forward in strengthening Europe's resilience against these sophisticated psychological attacks, thereby protecting its citizens and institutions in an increasingly interconnected world.

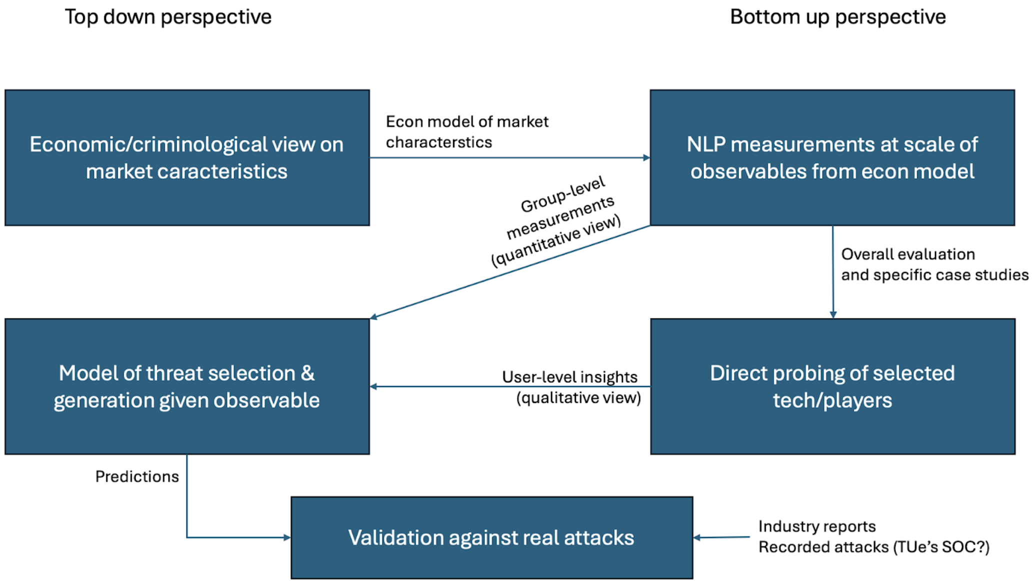


Figure 1 The envisioned approach to identify trends of emerging threats and attacker patterns.

Observing / Modeling Attacker Movements

Funding DoD (Department of Defense), GCHQ (Government Communications Headquarters), EU, DFG (German National Research Foundation), Cyber Insurances

Scope Creating threat actor ecosystem by 1) identifying trends of emerging (credible) threats and patterns of attacker interest in different attack capabilities made available in the darknet. 2) Characterize specific and emergent criminal convergence spaces from internet forums/anonymous markets to Telegram/Discord channels. 3) Quantify/qualify the effect of the appearance of an offensive capability in one of those venues for realizing an attack.

Pitch Social engineering attacks follow a pattern and defined phases (see 4.2). The pre-attack phases involve OSINT, target selection, and attack planning. This project aims to model the attacker patterns to predict social engineering attacks even before happening (i.e., prevention instead of mitigation). The overarching goal is developing an investigation infrastructure that triggers based on detecting attack patterns.

As shown in Figure 1 the approach would cover different aspects. Define economic and criminological theoretical underpinnings to identify measurables within forum and telegram/discord channels to characterize communities regarding the type of support (e.g., moral hazard/adverse selection mitigation mechanisms) they provide to criminal activities. NLP topic analysis will be used to identify both discussion topics within communities and user perception/feedback related to attack technology (sentiment analysis) to characterize user interactions at scale. Language and slang challenges must be addressed, especially on less verbose channels like instant messaging platforms. Active probing with direct interaction (either as potential customers/providers) or face-to-face interviews (remote setting) with offenders/perspective offenders to investigate underlying decision mechanisms/factors (e.g., why joining community x/choosing product y to do offense x rather than product y'). Develop a model of threat selection. Relate model predictions to high-level trends in emerging attack tech with known incidents and see if there is a credible link between what these markets enable and what attackers do.

Based on this *Threat Actor Ecosystem*, attacker movement would be modeled by:

1. Identifying patterns of attacker actions within network monitoring data.
2. Coding techniques on network packets/sequencing and MITRE ATT&CK mapping to evaluate qualitatively attack processes.
3. Building the conceptual model from that understanding.
4. Developing a data model for detecting those patterns from network data for scalability.
5. Evaluating the correlation of historical trends in those attack patterns with trends from our approach.

Live Threat Detection and Intervention (e.g., vishing)

Funding NSF (National Research Foundation), DFG (German National Research Foundation), Security companies

Scope Social Engineer's vishing dataset or self-data collection; project might include other modalities (keystroke dynamics, etc.)

Pitch An increasingly popular form of social engineering attacks is vishing (voice phishing). Little effective means for protection exist today against such attacks beyond raising awareness in cyber education. At the same time, the human voice holds rich information about (1) the current user state (i.e., whether they are stressed and what their current level of awareness is) as well as (2) techniques in use to social engineer somebody (firm voice to sound authoritative, pleading tone to beg for help, etc.). This project proposes to design, implement, and evaluate in-situ interventions protecting users from falling for vishing, that is, mechanisms that are capable of detecting in real-time if a caller is trying to social engineer somebody or if the callee is being socially engineered and provide active guidance as to how the legitimacy of a call can be verified.
The project will address the following objectives:
(1) building predictive models based on real vishing data, allowing common manipulation strategies and callee reactions to be detected;
(2) designing, implementing, and evaluating interventions to assist end users during vishing calls;
(3) assessing social and ethical implications of vishing mitigation technologies and strategies.

Using AI to support sharing content

Funding Social Media Platforms, Security Companies, and Research Foundations.

Pitch Social network users share personal information online that might be misused in several ways incl. social engineering. In this project, we want to investigate the usage AI-based support to inspect the information users want to share for (a) identifying content that could be misused or does not match the user's privacy needs, (b) educating the user on sharing consequences and (c) helping the users avoiding to share such information in the future, and (d) modify the content to mitigate the probability of misuse.

Using AI to design personalized support to mitigate vulnerabilities toward social engineering attack

Funding Research Foundations and Cyber Security Companies.

Pitch This funding proposal seeks support for an innovative project to leverage artificial intelligence (AI) to design personalized support systems that effectively mitigate vulnerabilities towards social engineering attacks. The proposed initiative recognizes the multifaceted nature of susceptibility, focusing on personability, age, cultural differences, and accessibility needs, among others, as critical dimensions to tailor interventions. The advent of sophisticated social engineering attacks demands a nuanced approach that adapts to individual characteristics. Our project will employ advanced AI algorithms to analyze and understand diverse user profiles to generate tailored responses and interventions resonating with users personally and build resilience to manipulation attempts.

The following ideas were identified but not further discussed:

- Utilizing AI for routine building and providing support
- Plugins and feature highlighting
- Designing social engineering interventions
- AI-based validation
- Self-assessment and self-reflection to mitigate vulnerabilities
- Design targeted support for vulnerable groups
- Understanding routines to mitigate vulnerabilities
- Characteristics of targets and social engineers

5 Report Summary

This report documents the outcomes of a three-day seminar that brought together experts from academia, industry, and authorities to address the escalating threats posed by social engineering in the digital age. The seminar aimed to develop a common understanding of social engineering, identify grand challenges, work on a research agenda, and foster collaboration in addressing social engineering vulnerabilities. Key themes included the professionalization of cyber attacks, the proliferation of sensors in everyday life as an opportunity for developing protection solutions, and the need for interdisciplinary collaboration to address evolving social engineering threats.

The seminar featured various sessions, including keynote speeches, group activities, and breakout discussions. Key discussions revolved around the fundamentals of social engineering, attack vectors, and the application of modern sensors to assess user states affected by social engineering. The participants also explored the ethical boundaries and obligations when using social engineering techniques for positive individual and social outcomes.

Furthermore, the seminar identified specific research areas and questions to be addressed through joint research projects and initiatives. Research areas included the evaluation of solutions, finding and supporting routines against attacks, and using datasets to understand threat actor communication and traceable transactions on the dark web.

The seminar outcomes underscore the importance of collaborative efforts between researchers and practitioners in fortifying against evolving social engineering threats. The insights gained from the seminar lay the foundation for future research and initiatives in addressing psychological manipulation in the digital age, including using modern sensors, ethical considerations in social engineering, and the development of protective measures against social engineering attacks.

Overall, the seminar provided a platform for interdisciplinary collaboration, fostering a deeper understanding of social engineering and its cognitive vulnerabilities. The identified grand challenges and proposed research projects highlight the significance of collaborative efforts in addressing the emerging threats posed by social engineering in the digital realm. The seminar outcomes provide valuable insights and potential research directions for fortifying against psychological manipulation and cyber threats.

6 Reading List

References

1 Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online." *ACM Comput. Surv.* 50, no. 3 (May 2018): 44. https://doi.org/10.1145/3054926

2 Florian Alt, Mariam Hassib, and Verena Distler. "Human-centered Behavioral and Physiological Security." In *Proceedings of the 2023 New Security Paradigms Workshop (NSPW '23)*, 2023: 48–61. https://doi.org/10.1145/3633500.3633504

3 Nattapat Boonprakong, Xiuge Chen, Catherine Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 27, 1–19. https://doi.org/10.1145/3544548.3580917

4 Lina Brunken, Annalina Buckmann, Jonas Hielscher, and M. Angela Sasse. (2023). "To Do This Properly, You Need More Resources": The Hidden Costs of Introducing Simulated Phishing Campaigns.

5 J-W. Bullée, and M. Junger. "How effective are social engineering interventions? A meta-analysis." *Information and Computer Security* 28, no. 5 (2020): 801–830. https://doi.org/10.1108/ICS-07-2019-0078

6 J.-W. Bullée, L. Montoya, W. Pieters, M. Junger, P. Hartel. "On the anatomy of social engineering attacks – A literature-based dissection of successful attacks". J Investig Psychol Offender Profil. 2018; 15: 20–45. https://doi.org/10.1002/jip.1482

7 Pavlo Burda, Luca Allodi, and Nicola Zannone. "Cognition in Social Engineering Empirical Research: a Systematic Literature Review." 2023. https://doi.org/10.1145/3635149

8 Felix Dietrich, Pascal Knierim, Yomna Abdelrahman, Ahmed Shams, Ken Pfeuffer, Mariam Hassib, and Florian Alt. "The Influence of Context on Response to Spear-Phishing Attacks: an In-Situ Deception Study." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, 2023: 619. https://doi.org/10.1145/3544548.3581170

9 Nina Gerber and Karola Marky. 2022. The nerd factor: the potential of S&P adepts to serve as a social resource in the user's quest for more secure and privacy-preserving behavior. In Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security (SOUPS'22). USENIX Association, USA, Article 4, 57–76.

10 Martina Angela Sasse and Iacovos Kirlappos, "Security Education against Phishing: A Modest Proposal for a Major Rethink" in IEEE Security and Privacy, vol. 10, no. 02, pp. 24-32, 2012. https://doi.ieeecomputersociety.org/10.1109/MSP.2011.179

11 Karola Marky, Martin Schmitz, Verena Zimmermann, Martin Herbers, Kai Kunze, and Max Mühlhäuser. 2020. 3D-Auth: Two-Factor Authentication with Personalized 3D-Printed Items. In Proceedings of the 2020 CHI Conference on Human Factors in Computing

Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376189

12 Karola Marky, Shaun Macdonald, Yasmeen Abdrabou and Mohamed Khamis. 2023. In the Quest to Protect Users from Side-Channel Attacks – A User-Centred Design Space to Mitigate Thermal Attacks on Public Payment Terminals. 32nd USENIX Security Symposium (USENIX Security 23). https://www.usenix.org/conference/usenixsecurity23/presentation/marky

13 Florian Mathis, John H. Williamson, Kami Vaniea, and Mohamed Khamis. 2021. Fast and Secure Authentication in Virtual Reality Using Coordinated 3D Manipulation and Pointing. ACM Trans. Comput.-Hum. Interact. 28, 1, Article 6 (February 2021), 44 pages. https://doi.org/10.1145/3428121

14 Peter Mayer, Yixin Zou, Byron M. Lowens, Hunter A. Dyer, Khue Le, Florian Schaub, and Adam J. Aviv. 2023. Awareness, Intention, (In)Action: Individuals' Reactions to Data Breaches. ACM Trans. Comput.-Hum. Interact. 30, 5, Article 77 (October 2023), 53 pages. https://doi.org/10.1145/3589958

15 Simone Ooms, Minha Lee, Pablo Cesar, and Abdallah El Ali. 2023. FeelTheNews: Augmenting Affective Perceptions of News Videos with Thermal and Vibrotactile Stimulation. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 137, 1–8. https://doi.org/10.1145/3544549.3585638

16 Justin Petelka, Yixin Zou, and Florian Schaub. 2019. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 518, 1–15. https://doi.org/10.1145/3290605.3300748

17 Christina Schneegass, Thomas Kosch, Andrea Baumann, Marius Rusu, Mariam Hassib, and Heinrich Hussmann. 2020. BrainCoDe: Electroencephalography-based Comprehension Detection during Reading and Listening. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376707

18 Steeven Villa, Thomas Kosch, Felix Grelka, Albrecht Schmidt, and Robin Welsch. 2023. The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior. Comput. Hum. Behav. 146, C (Sep 2023). https://doi.org/10.1016/j.chb.2023.107787

19 Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. 2020. Analysing Simulated Phishing Campaigns for Staff. In Computer Security: ESORICS 2020 International Workshops, DETIPS, DeSECSys, MPS, and SPOSE, Guildford, UK, September 17–18, 2020, Revised Selected Papers. Springer-Verlag, Berlin, Heidelberg, 312–328. https://doi.org/10.1007/978-3-030-66504-3_19

20 Yixin Zou, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. 2020. Examining the Adoption and Abandonment of Security, Privacy, and Identity Theft Protection Practices. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376570

Participants

- Yomna Abdelrahman
European Universities in Egypt – Cairo, EG
- Luca Allodi
TU Eindhoven, NL
- Florian Alt
University of the Bundeswehr Munich, DE
- Nathan Berry
Nexus – Leeds, GB
- Jan-Willem Bullee
University of Twente, NL
- Mary D'Angelo
Searchlight Cyber – Washington, DC, US
- Felix Dietz
University of the Bundeswehr Munich, DE
- Tilman Dingler
The University of Melbourne, AU
- Verena Distler
University of the Bundeswehr Munich, DE
- Abdallah El Ali
CWI – Amsterdam, NL
- Jerry Färdigs
Swedish Armed Forces – Uppsala, SE
- Ann Fernström
Swedish Armed Forces – Uppsala, SE
- Matteo Große-Kampmann
AWARE7 GmbH – Gelsenkirchen, DE
- Christopher Hadnagy
Social-Engineer – Orlando, US
- Mohamed Khamis
University of Glasgow, GB
- Claude Kirchner
CCNE – Paris, FR & INRIA – Rocquencourt, FR
- Thomas Kosch
HU Berlin, DE
- Karola Marky
Ruhr-Universität Bochum, DE
- Abbie Maroño
Social-Engineer – Orlando, US
- Alexander Nussbaum
University of the Bundeswehr Munich, DE
- Alia Saad
Universität Duisburg-Essen, DE
- Martina Angela Sasse
Ruhr-Universität Bochum, DE
- Florian Schaub
University of Michigan – Ann Arbor, US
- Christina Schneegass
TU Delft, NL

Report from Dagstuhl Seminar 23471

The Next Generation of Deduction Systems: From Composition to Compositionality

Maria Paola Bonacina[*1], **Pascal Fontaine**[*2], **Cláudia Nalon**[*3], **Claudia Schon**[*4], **and Martin Desharnais**[†5]

1 **University of Verona, IT.** mariapaola.bonacina@univr.it
2 **University of Liège, BE.** pascal.fontaine@uliege.be
3 **University of Brasília, BR.** nalon@unb.br
4 **Hochschule Trier, DE.** c.schon@hochschule-trier.de
5 **Max-Planck-Institut für Informatik Saarbrücken, DE.** desharnais@mpi-inf.mpg.de

Abstract

Deduction systems are computer procedures that employ inference or transition rules, search strategies, and multiple supporting algorithms, to solve problems by logico-deductive reasoning. They are at the heart of *SAT/SMT solvers*, *theorem provers*, and *proof assistants*. The wide range of successful applications of these tools shows how logico-deductive reasoning is well-suited for machines. Nonetheless, *satisfiability* and *validity* are difficult problems, and applications require reasoners to handle *large and heterogeneous knowledge bases*, and to generate *proofs* and *models* of increasing size and diversity. Thus, a vast array of techniques was developed, leading to what was identified during the seminar as a *crisis of growth*. This crisis manifests itself also as a software crisis, called *automated reasoning software crisis* at the seminar. Many deduction systems remain prototypes, while relatively few established systems resort to assemble techniques into *portfolios* that are useful for experiments, but do not lead to breakthroughs.

In order to address this *crisis of growth*, the Dagstuh Seminar "The Next Generation of Deduction Systems: From Composition to Compositionality" (23471) focused on the key concept of *composition*, that is, a combination where properties of the components are preserved. Composition applies to all building blocks of deduction: *rule systems*, *strategies*, *proofs*, and *models*. All these instances of compositions were discussed during the seminar, including for example composition of *instance-based and superposition-based inference systems*, and composition of modules towards *proof production* in SMT solvers. Other kinds of composition analyzed during the seminar include the composition of *reasoning and learning*, and the composition of *reasoning systems* and *knowledge systems*. Indeed, reasoners *learn* within and across derivations, while for applications, from *verification* to *robotics*, provers and solvers need to work with other knowledge-based components.

In order to address the *automated reasoning software crisis*, the seminar elaborated the concept of *compositionality*, as the engineering counterpart of what is composition at the theory and design levels. The seminar clearly identified *modularity* as the first step towards compositionality, proposing to decompose existing systems into *libraries of modules* that can be recomposed in new systems. The ensuing discussion led to the distinction between automated reasoners that are *industry powertools* and automated reasoners that are *pedagogical tools*. At the societal level, this distinction is important to counter the phenomenon whereby new students are either discouraged by the impossibility of competing with industry powertools, or induced to join only those research groups that work on industry powertools. In summary, the seminar fully succeeded in promoting the exchange of ideas and suggestions for future work.

Seminar November 19–24, 2023 – https://www.dagstuhl.de/23471
2012 ACM Subject Classification Theory of computation → Automated reasoning

[*] Editor / Organizer
[†] Editorial Assistant / Collector

The Next Generation of Deduction Systems: From Composition to Compositionality, *Dagstuhl Reports*, Vol. 13, Issue 11, pp. 130–150
Editors: Maria Paola Bonacina, Pascal Fontaine, Cláudia Nalon, and Claudia Schon
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Keywords and phrases artificial intelligence, automated reasoning, compositionality, deduction, logic
Digital Object Identifier 10.4230/DagRep.13.11.130

1 Executive Summary

Maria Paola Bonacina
Pascal Fontaine
Cláudia Nalon
Claudia Schon

This report contains the program and outcomes of the Dagstuhl Seminar 23471 on *The Next Generation of Deduction Systems: From Composition to Compositionality* that was held at Schloss Dagstuhl, Leibniz Center for Informatics, during November 19–24, 2023. It was the fifteenth in a series of Dagstuhl Deduction seminars held biennially since 1993.

The motivation for this seminar was threefold:

1. Automated reasoning tools, including *SAT solvers*, *SMT solvers*, *theorem provers*, and *proof assistants*, are widely applied in fields as diverse as *analysis/verification/synthesis of systems*, *programming language design*, *knowledge engineering*, *computer mathematics*, *natural language processing*, and *robotics*. However, *satisfiability* and *validity* remain *fundamentally difficult computational problems*, so that reasoners may run out of time or memory returning "don't know" or may demand too much human labor.
2. After the low-hanging fruits have been picked, the *formalization* of problems require logics, formulas, theories, and knowledge bases that are increasingly complex, large, and heterogeneous. The *size* and *diversity* of the *proofs* and *models*, that reasoners produce to support their answers, increase accordingly.
3. Deduction offers a vast array of techniques, but many implementations of new techniques remain short-lived prototypes, and the transfer of the successful ones into more stable systems is uncertain. Relatively few systems gather most of the resources, but over time they may become too big, monolithic, and unwieldy for further development, or resort to assemble techniques into portfolios. A portfolio allows one to experiment and may win competitions, but it hardly leads to a conceptual synthesis and hence a breakthrough.

The Dagstuhl Seminar on *The Next Generation of Deduction Systems: From Composition to Compositionality* addressed these issues by challenging participants to reflect around the ideas of *composition* and *compositionality*.

A *composition* is a combination such that properties of the components (e.g., *soundness*, *completeness*, *termination*, *model-construction*) are *preserved*. Since different inference systems have different strengths, their composition is essential to meet Challenge (1). For example, the seminar participants presented and discussed research about the composition of equality reasoning by *superposition* with *instance-based* (e.g., *Inst-gen* – "Instance Generation") or *model-based* (e.g., *SCL* – "Simple Clause Learning") inference systems.

A major cause of "don't know" answers in *satisfiability modulo theories* (SMT) is the fact that most decision procedures inside SMT solvers are for quantifier-free fragments of theories, whereas applications require handling quantified formulas. Thus, the seminar addressed the fundamental problem of *composing quantifier reasoning and theory reasoning*. For example, the *QSMA* algorithm, where *QSMA* stands for "Quantified Satisfiability Modulo Assignment," offers a novel solution for quantifier reasoning in a complete theory (e.g., arithmetic).

Historically, *proof generation* was deemed unproblematic in automated theorem provers, whereas *model generation* was deemed unproblematic in SMT solving. This is why recent research has focused on proofs in SMT and models in first-order theorem proving. The seminar reflected these trends. Several talks presented advanced research on *proof production* in SMT, involving *composition of proofs*, both within the SMT solver, as in composition of proofs from different theories, and at the interface of the SMT solver (e.g., CVC5) with a proof assistant (e.g., Lean, Isabelle/HOL). At the next abstraction level, the seminar analyzed these issues in *logical frameworks* (e.g., *Hybrid*, *Dedukti*), where proofs from different proof assistants may be verified, exchanged, translated, and hence re-verified. Work on the representation and composition of first-order models in libraries of problems for first-order theorem provers (e.g., TPTP) is also gaining momentum, and the seminar offered an excellent discussion forum, since several developers of theorem provers were attending.

The drive to improve the search capabilities of deduction procedures in order to meet Challenge (1) leads also to the composition of *reasoning* and *learning*, while Challenge (2) leads to the composition of *reasoning systems* and *knowledge systems*. Learning is a native capability of automated reasoners, as in *lemma learning*. SAT/SMT solvers and theorem provers *learn within a derivation* by learning lemmas to reduce the search space by avoiding repeated work. Reasoners also *learn across derivations* by applying *machine learning* to learn from a very high number of derivations which *strategies* or *tactics* to select for an input problem with certain *features*. The composition of reasoning and learning was discussed at the seminar in *SAT solving*, and in *resolution-based* first-order theorem proving, where the prover is interfaced with an ontology-based knowledge system (e.g., Adimen SUMO).

The sentiment that emerged at the seminar is that approaches based on *composition* will contribute to meet Challenge (3), by endowing deduction systems with *compositionality*, towards going beyond portfolios. The participants discussed the *crisis of growth* that the field is facing, given the rise of so many rule systems, strategies, and techniques. Since it is a crisis of *growth*, the field will emerge from it even stronger. For this to happen, however, it is key to address the issues that make it difficult to transfer new ideas into stable and useable deduction systems. The existing dichotomy, between short-lived prototypes and powerful, but big, monolithic, unwieldy systems, was discussed as an *automated reasoning software crisis*. The need for *modularity* was recognized, and a distinction between *industry powertools* and *pedagogical platforms* was outlined. The latter will have to give up on a unique programming language and programming style, as well as on award-winning efficiency, but will facilitate the entrance of new students, currently discouraged by the impossibility of competing with established tools. Thanks to such platforms, the building of new systems will be less expensive in terms of human time and labor. The risk of new ideas being forgotten without having been properly implemented and tested will be reduced.

The atmosphere throughout the seminar was excellent. For example, a participant told one of the organizers that this seminar motivated them and rekindled their enthusiasm for automated deduction research. An outing – an excursion to Bernkastel-Kues followed by a social dinner in a nearby village – also contributed to establishing a relaxed, friendly atmosphere, conducive to new or strengthtened collaborations.

The bottom-up style of the Dagstuhl experience was preserved, thanks to a flexible program that allowed the participants to volunteer topics and talks throughout the gathering. This seminar maintained a feature that was introduced in the 2021 edition, namely the possibility of giving a tutorial using two time slots rather than one. Altogether, *five tutorials* were given on topics ranging from *proofs in SMT*, *reasoning with quantifiers in SMT*, *composition of reasoning and neuro-symbolic methods*, and *model-based reasoning*.

The following section contains the abstracts for most of the talks and tutorials listed in alphabetical order.

2 Table of Contents

3 Overview of Talks

3.1 Combining Proofs for Description Logic and Concrete Domain Reasoning

Franz Baader (TU Dresden, DE)

Joint work of Christian Alrabbaa, Franz Baader, Stefan Borgwardt, Patrick Koopmann, Alisa Kovtunova
Main reference Christian Alrabbaa, Franz Baader, Stefan Borgwardt, Patrick Koopmann, Alisa Kovtunova: "Combining Proofs for Description Logic and Concrete Domain Reasoning", in Proc. of the Rules and Reasoning – 7th International Joint Conference, RuleML+RR 2023, Oslo, Norway, September 18-20, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14244, pp. 54–69, Springer, 2023.
URL https://doi.org//10.1007/978-3-031-45072-3_4

Logic-based approaches to AI have the advantage that their behavior can in principle be explained with the help of proofs of the computed consequences in an appropriate calculus. To benefit from this in practice, considerable work beyond the implementation of a reasoning system is needed to be able to compute proofs that are appropriate for explanation purposes. For ontologies based on Description Logic (DL), we have put this advantage into practice by showing how proofs for consequences derived by DL reasoners can be computed and displayed in a user-friendly way. However, these methods are insufficient in applications where also numerical reasoning is relevant. The present paper considers proofs for DLs extended with concrete domains (CDs) based on the rational numbers, which leave reasoning tractable if integrated into the lightweight DL $\mathcal{EL}_\bot$. Since no implemented DL reasoner supports these CDs, we first develop reasoning procedures for them, and show how they can be combined with reasoning approaches for pure DLs, both for $\mathcal{EL}_\bot$ and the more expressive DL $\mathcal{ALC}$. These procedures are designed such that it is easy to extract proofs from them. We show how the extracted CD proofs can be combined with proofs on the DL side into integrated proofs that explain both the DL and the CD reasoning. We have implemented our reasoning and proof extraction approaches for DLs with concrete domains and have evaluated them on several self-created benchmarks.

3.2 SMT Proof Production and Integration with the Lean Theorem Prover

Haniel Barbosa (Federal University of Minas Gerais-Belo Horizonte, BR)

Joint work of Haniel Barbosa, Tomaz Gomes Mascarenhas, Bruno Andreotti, Andrew Reynolds, Gereon Kremer, Hanna Lachnitt, Aina Niemetz, Andres Nötzli, Alex Ozdemir, Mathias Preiner, Arjun Viswanathan, Scott Viteri, Yoni Zohar, Cesare Tinelli, Clark Barrett
Main reference Haniel Barbosa, Andrew Reynolds, Gereon Kremer, Hanna Lachnitt, Aina Niemetz, Andres Nötzli, Alex Ozdemir, Mathias Preiner, Arjun Viswanathan, Scott Viteri, Yoni Zohar, Cesare Tinelli, Clark W. Barrett: "Flexible Proof Production in an Industrial-Strength SMT Solver", in Proc. of the Automated Reasoning – 11th International Joint Conference, IJCAR 2022, Haifa, Israel, August 8-10, 2022, Proceedings, Lecture Notes in Computer Science, Vol. 13385, pp. 15–35, Springer, 2022.
URL https://doi.org//10.1007/978-3-031-10769-6_3

SMT solvers can be hard to trust, since it generally means assuming their large and complex codebases do not contain bugs leading to wrong results. Machine-checkable certificates, via proofs of the logical reasoning the solver has performed, address this issue by decoupling confidence in the results from the solver's implementation. In this talk we will describe extensive proof infrastructure of the state-of-the-art SMT solver cvc5, which has enabled the production of proofs in a number of complex domains. We will also show ongoing work towards integrating these proofs into the proof assistant Lean, thus enabling its composition with SMT solvers in a trusted way.

3.3 The QSMA algorithm

Maria Paola Bonacina (University of Verona, IT)

Joint work of Maria Paola Bonacina, Stéphane Graham-Lengrand, Christophe Vauthier
Main reference Maria Paola Bonacina, Stéphane Graham-Lengrand, Christophe Vauthier: "QSMA: A New Algorithm for Quantified Satisfiability Modulo Theory and Assignment", in Proc. of the Automated Deduction – CADE 29 – 29th International Conference on Automated Deduction, Rome, Italy, July 1-4, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14132, pp. 78–95, Springer, 2023.
URL https://doi.org//10.1007/978-3-031-38499-8_5

Automated theorem provers (ATP) for first-order or higher-order logic and solvers for satisfiability modulo theories (SMT) exhibit impressive power and sophistication. ATP systems reason well about formulas with free symbols and universally quantified variables, removing existential quantifiers by Skolemization. SMT solvers reason well about formulas with free or existentially quantified variables and symbols defined by a theory. However, formulas from key applications involve both arbitrary quantification and defined symbols. The successful composition of quantifier and theory reasoning is a major objective for the next generation of deduction systems. QSMA is a new algorithm for quantifiers in SMT. QSMA stands for Quantified Satisfiability Modulo theory and Assignment. Currently, QSMA works for one theory with unique interpretation of symbols (e.g., arithmetic), so that models differ only in the assignment to free variables. QSMA accepts arbitrary formulas: the quantifiers may alternate and occur in arbitrary positions, as not even prenex normal form is required. After turning universal quantifiers into existential ones by double negation, QSMA performs a recursive descent over the tree structure of the formula, peeling off quantifiers and instantiating variables. Thus, each call works modulo assignment. By building under- and over- approximations of the formula, QSMA zooms in on a model or finds that none exists. The YicesQS solver implements QSMA on top of the Yices 2 solver and exhibits excellent performance in arithmetic. Composing QSMA within the CDSAT framework for conflict-driven satisfiability in a union of theories is the next challenge.

(QSMA is joint work with Stéphane Graham-Lengrand and Christophe Vauthier. CDSAT is joint work with Stéphane Graham-Lengrand and Natarajan Shankar. Stéphane Graham-Lengrand is the author of YicesQS. Bruno Dutertre and Dejan Jovanović are the authors of Yices 2.)

3.4 An Isabelle/HOL Formalization of the SCL(FOL) Calculus

Martin Desharnais (Max-Planck-Institut für Informatik Saarbrücken, DE)

Joint work of Martin Bromberger, Martin Desharnais, Christoph Weidenbach
Main reference Martin Bromberger, Martin Desharnais, Christoph Weidenbach: "An Isabelle/HOL Formalization of the SCL(FOL) Calculus", in Proc. of the Automated Deduction – CADE 29 – 29th International Conference on Automated Deduction, Rome, Italy, July 1-4, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14132, pp. 116–133, Springer, 2023.
URL https://doi.org//10.1007/978-3-031-38499-8_7

We present an Isabelle/HOL formalization of SCL(FOL): Simple Clause Learning for first-order logic without equality. The main results are formal proofs of soundness, non-redundancy of learned clauses, termination, and refutational completeness. Compared to the unformalized version, the formalized calculus is simpler, a number of results could be generalized, and the non-redundancy property strengthened. We found one bug in a previously published version of the SCL Backtrack rule. Compared to related formalizatons, we introduce a new technique for showing termination based on non-redundant clause learning.

3.5 Compositionality from Temporal Logics to Verification for Autonomous Robot Systems

Clare Dixon (University of Manchester, GB)

This talk was split into two parts: firstly relating to a resolution based calculus and its implementation for propositional linear-time temporal logic and secondly relating to experiences with verification for autonomous robots. With respect to the former I discussed a resolution calculus for proposition linear-time temporal logic and showed how some of the resolution rules could be implemented by calls to a first order logic prover (composition). Secondly I discussed more recent work towards verification for robots (compositionality). Two approaches to verification were mentioned: heterogeneous verification and corroborative verification. With heterogeneous verification we need the robot system being considered to be split into modular subcomponents. On each subcomponent we apply the most suitable verification (including both formal or non-formal verification) for that subsystem for example model checking, theorem proving, software testing, simulation based testing, real robot experiments etc. For each component the assumptions on inputs made on the system eventually must be shown to guarantee required outputs. Ongoing work from colleagues involves how to compose such results to get an overall confidence in the system. Corroborative verification involves applying different verification types to a (sub) system and utilising the outputs to improve the verification models and properties for the other verification types increasing the confidence in the systems.

3.6 (Re)Verification of Proofs with Coq or Dedukti

Catherine Dubois (ENSIIE – Evry, FR)

Joint work of Catherine Dubois, Chantal Keller
Main reference Valentin Blot, Denis Cousineau, Enzo Crance, Louise Dubois de Prisque, Chantal Keller, Assia Mahboubi, Pierre Vial: "Compositional Pre-processing for Automated Reasoning in Dependent Type Theory", in Proc. of the 12th ACM SIGPLAN International Conference on Certified Programs and Proofs, CPP 2023, Boston, MA, USA, January 16-17, 2023, pp. 63–77, ACM, 2023.
URL https://doi.org//10.1145/3573105.3575676
Main reference Burak Ekici, Alain Mebsout, Cesare Tinelli, Chantal Keller, Guy Katz, Andrew Reynolds, Clark W. Barrett: "SMTCoq: A Plug-In for Integrating SMT Solvers into Coq", in Proc. of the Computer Aided Verification – 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 10427, pp. 126–133, Springer, 2017.
URL https://doi.org//10.1007/978-3-319-63390-9_7
Main reference Frédéric Blanqui, Gilles Dowek, Émilie Grienenberger, Gabriel Hondet, François Thiré: "A modular construction of type theories", Log. Methods Comput. Sci., Vol. 19(1), 2023.
URL https://doi.org//10.46298/LMCS-19(1:12)2023
Main reference Guillaume Burel, Guillaume Bury, Raphaël Cauderlier, David Delahaye, Pierre Halmagrand, Olivier Hermant: "First-Order Automated Reasoning with Theories: When Deduction Modulo Theory Meets Practice", J. Autom. Reason., Vol. 64(6), pp. 1001–1050, 2020.
URL https://doi.org//10.1007/S10817-019-09533-Z
Main reference Mohamed Yacine El Haddad, Guillaume Burel, Frédéric Blanqui: "EKSTRAKTO A tool to reconstruct Dedukti proofs from TSTP files (extended abstract)", in Proc. of the Proceedings Sixth Workshop on Proof eXchange for Theorem Proving, PxTP 2019, Natal, Brazil, August 26, 2019, EPTCS, Vol. 301, pp. 27–35, 2019.
URL https://doi.org//10.4204/EPTCS.301.5

Verifying or cross-verifying proofs improves the confidence we have in proofs. The talk focusses on the use of Coq or Dedukti as proof checkers. We first give a quick overview of SMTCoq and the recent tactic sniper that allows for more automation when a Coq first order goal is discharged using a SAT/SMT solver. Then we briefly introduce the Dedukti logical framework. The last part of the talk quickly presents the proof tools Zenon Modulo, iProverModulo, Archsat, and Ekstrakto. The three first ones directly produce Dedukti proofs that can be checked by the Dedukti checker. The latter reconstructs a Dedukti proof from a proof trace by reproving each step using a Dedukti producing tool and combining the proofs of the steps to get a proof of the original formula. Finally we point out 2 projects: BWare and ICSPA. The first one aimed at developing a mechanized framework for automated verification of AtelierB proof obligations where Zenon Modulo and iProvermodulo were used. ICSPA is a project in progress where the objectives are to improve confidence in the proofs realized in the context of B/Event-B and TLA+ by formally and independently verifying these proofs and also enable sharing and reusing proofs and models between B/Event-B and TLA+ using lambda-PI calculus modulo theory and Dedukti.

3.7 Formal Verification at CLEARSY : Needs and Prospects

David Déharbe (CLEARSY – Aix-en-Provence, FR)

The talk first briefly presents CLEARSY, a French SME created in the early 2000s that literally has formal methods in its DNA, as it was created to promote the B method and to distribute and maintain Atelier B, the tooling of the B method. The B method, is a rigorous, logic-based framework to design correct-by-construction software components. Invented by

J.R. Abrial, its first industrial application has been software that safeguards a fully automatic metro line in Paris, France. The talk gives some technical details on the B method, so that it should be evident that a robust, reliable, efficient and versatile automatic proof support is essential to make this method even more attractive, by reducing the burden to interact with a proof assistant to discharge proof obligations (POs), and therefore to make it more competitive. Formal verification support in Atelier B has historically relied on custom automatic provers (pp and pr) and proof assistant (pri). The talk presents how third-party provers may now be used in Atelier B, through extension points called proof mechanisms. A proof mechanism is here a combination of tool chains made of an external prover, a translator that encodes the logic of B to that of the prover, and an interpreter for the prover's output. Several such tool chains may be applied to the same PO to increase proof coverage, or trust in the result, or both. Finally, the talk presents novel ideas to improve such proof mechanisms by taking advantage of the capability of ATP systems and SMT solvers to produce proofs and so-called unsat cores. Indeed an unsat core identifies a subset of the (usually very large number of) hypotheses that is sufficient to prove the goal is valid. We can then use an unsat core to build a reduced PO that may then be more easily processed by another tool chain. We thus expect to achieve a much higher coverage for each tool chain and eventually improved confidence and efficiency in using third-party provers in Atelier B.

3.8 Reasoning with Structured Contexts of Assumptions

Amy Felty (University of Ottawa, CA)

Joint work of Amy Felty, Mohamed Yousri Mahmoud, Alberto Momigliano, Brigitte Pientka

We present past and current work on adding support for reasoning on open terms with structured contexts of assumptions in the Hybrid logical framework (LF). Hybrid is implemented in Coq and is designed to support the use of higher-order abstract syntax (HOAS), also called lambda-tree syntax, for representing and reasoning about formal systems such as logics and programming languages. In previous work, we considered a large class of intuitionistic LFs supporting HOAS, and introduced a common infrastructure and general language for structuring such reasoning on open terms with structued contexts, along with some benchmarks. Our recent work has also included large case studies in a linear logic version of Hybrid.

In this talk, we discuss combining and extending our past work in these directions. In particular, we present a variety of examples specific to Hybrid and our case studies, both intuitionistic and linear, and discuss our planned work on extending the general infrastructure and language designed for intuitionistic LFs to the setting of linear LFs. We also discuss automating the generation of lemmas and proofs in both the intuitionistic and linear settings.

References

1 Amy Felty, Alberto Momigliano, and Brigitte Pientka. Benchmarks for reasoning with syntax trees containing binders and contexts of assumptions. *Mathematical Structures in Computer Science*, 28:1507–1540, 2018.

2 Amy P. Felty and Alberto Momigliano. Hybrid: A definitional two-level approach to reasoning with higher-order abstract syntax. *Journal of Automated Reasoning*, 48(1):43–105, 2012.

3 Mohamed Yousri Mahmoud and Amy P. Felty. Formalization of metatheory of the quipper quantum programming language in a linear logic. *Journal of Automated Reasoning*, 63:967–1002, 2019.

3.9 On the need for a modular approach for automated reasoners

Pascal Fontaine (University of Liège, BE)

Joint work of Pascal Fontaine, Haniel Barbosa, Martin Bromberger, Sophie Tourret

In this short presentation, essentially meant to stimulate the discussion among participants, I exposed a very subjective view on the evolution of automated reasoning software, from many small one-person projects in the 90s to a few huge tools now. This poses a problem for the future of the field. I advocate a modular approach to software in our field, to enable reuse, for better distribution of the work, for students to more easily understand the tools by parts, and for better evaluation of parts of automated reasoning software. I briefly reported on my first experiment for a modular approach in SMT with modulariT.

3.10 Interpolation Properties for Array Theories: Positive and Negative Results

Silvio Ghilardi (University of Milan, IT)

Main reference Silvio Ghilardi, Alessandro Gianola, Deepak Kapur, Chiara Naso: "Interpolation Results for Arrays with Length and MaxDiff", ACM Trans. Comput. Log., Vol. 24(4), pp. 28:1–28:33, 2023.
URL https://doi.org//10.1145/3587161

In this talk, we first review basic correspondences between syntactic interpolation properties of a first order theory (quantifier-free interpolation property, general quantifier-free interpolation property, uniform quantifier-free interpolation property) and semantic features related to the class of its models (amalgamation, strong amalgamation, model completability). Then we shall analyze these notions for variants of McCarthy extensional theory of arrays. Whereas the basic theory does not have quantifier-free interpolation property, such property can be restored by adding it an extra symbol 'diff' skolemizing the extensionality axiom. General quantifier-free interpolation property also holds for this theory but not uniform quantifier-free interpolation property, as shown by an explicit counterexample. Since the semantic content of diff operation is rather underspecified, we strenghten the theory by asking diff(a,b) to return the maximum index where two arrays a,b differ (diff returns 0 if they are equal). We also add to a unary 'length' operation. We so end up in a theory still having quantifier-free interpolation, as witnessed by a hyerarchic polynomial reduction to general interpolation for linear arithmetics over indexes. General quantifier free interpolation property may fail, but can be re-gained by introducing constant arrays.

The second part of this talk comes from joint work with A. Gianola, D. Kapur, C. Naso [ACM-TOCL, October 2023]. The first part of the talk reviews old joint work with R. Bruttomesso and S. Ranise, adding to such old work some recent achievements.

3.11 Formal Verification at Certora

Antti Hyvärinen (Certora – Pregassona, CH)

Traditional finance is largely based on the assumption that human actors behave in a trustworthy manner. When this trust was misplaced, this has resulted in big losses for financial systems. Decentralized finance (DeFi) provides a solution by making financial protocols transparent and automated. As a result DeFi does not have the guardrails provided by humans, and catastrophic failures result from incorrect implementations. Certora's bounded model checking based tool helps finding faults in the protocols in an exhaustive way. In this talk I describe how a critical bug was found and fixed in a protocol design and how the tool helped in this process.

3.12 Improving SMT Solving via Incorporating More Techniques

Fuqi Jia (Chinese Academy of Sciences, CN)

Joint work of Fuqi Jia, Feifei Ma, Minghao Liu

In this talk, we would like to introduce some new approaches to solving the SMT problem, including: 1. A bit-blasting based algorithm for SMT(NIA) formulas; 2. A gradient-based algorithm for SMT(NRA) formulas; 3. SMT solving under probability distribution. These works explored the advancement of four components of SMT solving: Search Space Allocation, Variable Order Selection, Model or Partial Model Generation, and Value Decision.

3.13 Higher-order constraint term rewriting

Cynthia Kop (Radboud University Nijmegen, NL)

Joint work of Cynthia Kop, Liye Guo

Logically Constrained Term Rewriting Systems offer a way to couple traditional reasoning on term rewriting systems with SMT reasoning (and tools). This allows them, in turn, to be used for program analysis in a more natural way than pure rewriting (and in different ways than pure SMT). But to model functional languages naturally, we should ideally combine *higher-order* term rewriting systems with SMT. In this presentation, I will discuss the choices to be made for that goal.

3.14 Reconstruction of cvc5 Proofs in Isabelle/HOL

Hanna Lachnitt (Stanford University, US)

Joint work of Hanna Lachnitt, Mathias Fleury, Andrew Reynolds, Haniel Barbosa, Andres Noetzli, Leni Aniva, Clark Barrett, Cesare Tinelli

The proof assistant Isabelle/HOL can call external solvers to automate proof search, which is crucial for using it more effectively. In particular, statements containing bit-vectors are notoriously tedious to prove manually. cvc5 is an efficient satisfiability modulo theories (SMT) solver that is currently only indirectly used by Isabelle. The process of finding a proof inside of Isabelle with the information provided by cvc5 is slow and often fails. In this work we extend the integration between Isabelle and cvc5 so that a proof certificate from cvc5 is shared with Isabelle that can be reconstructed internally into native Isabelle/HOL proofs. We present our ongoing effort to reconstruct these proofs, including problems containing bit-vectors whose reconstruction in Isabelle is currently not supported by any other SMT solver. Modern SMT solvers implement hundreds of term rewriting rules. cvc5 is able to output fine-grained proofs using a separate database of rewrite rules written in the RARE language. We also present IsaRARE, a plugin for Isabelle, that translates such rules to lemmas in Isabelle that can then be used in the reconstruction process out of the box. Additionally, IsaRARE can be used as a verifier for rewrite rules. We evaluate our approach by verifying an extensive set of rewrite rules used by the cvc5 SMT solver.

3.15 Solving Reasoning Problems with Neuro-Symbolic Methods

Feifei Ma (Chinese Academy of Sciences – Beijing, CN) and Fuqi Jia (Chinese Academy of Sciences, CN)

Joint work of Feifei Ma, Fuqi Jia, Minghao Liu

Symbolism and connectionism are two fundamental paradigms for artificial intelligence. In the past decade, connectionism has revived in the name of deep learning, achieving great success in many areas. Recently, neuro-symbolic methods, aiming to bridge the gap between connectionism and symbolism, receive much attention. In this talk, we will introduce some of our initial efforts in this area, which can be classified into two categories: 1. The end-to-end approach where a neural network takes as input the reasoning task and directly outputs the result; 2. The composition of neural network and symbolic method, where a neural network provides assistance to the reasoning algorithm. The targeted reasoning problems include pseudo-Boolean constraint solving, MaxSAT and cylindrical algebraic decomposition.

3.16 A Compositional Proof System for Cylindrical Algebraic Decomposition

Jasper Nalbach (RWTH Aachen, DE)

Joint work of asper Nalbach, Erika Ábrahám, Philippe Specht, Christopher W. Brown, James H. Davenport, Matthew England
Main reference Jasper Nalbach, Erika Ábrahám, Philippe Specht, Christopher W. Brown, James H. Davenport, Matthew England: "Levelwise construction of a single cylindrical algebraic cell", CoRR, Vol. abs/2212.09309, 2022.
URL https://doi.org//10.48550/ARXIV.2212.09309

Cylindrical algebraic decomposition (CAD) is the only complete method implemented in Satisfiability-modulo-theories solvers for solving non-linear arithmetic. Due to its doubly exponential complexity, modern algorithms compute only parts of its projection operation, making solving some practical instances of NRA tractable. There is a variety of cases where savings in the projection are possible, and often there are multiple alternatives for the projection. To manage the maintainability of an algorithm when incorporating special cases, we developed a proof system for modern CAD-based SMT algorithms. This proof system is extensible, separates heuristic decisions (which projection to take) from the correctness of the projection and can be employed in different algorithms. Further, the proof system could be a step towards formal proofs for real algebra.

3.17 A Unified Proof System for Discrete Combinatorial Problems

Jakob Nordström (University of Copenhagen, DK & Lund University, SE)

URL https://gitlab.com/MIAOresearch/software/VeriPB

We give a brief overview of VeriPB, a proof system based on pseudo-Boolean reasoning with 0-1 integer linear inequalities that seems well suited to provide a unified proof logging method for discrete combinatorial problems. We have implemented VeriPB proof logging, together with efficient proof checking, for state-of-the-art solvers in Boolean satisfiable (SAT) solving, SAT-based optimization, graph solving, constraint programming, and a growing list of other combinatorial solving paradigms. We believe that ideas from VeriPB could be useful also in the context of mixed integer linear programming and satisfiability modulo theories (SMT) solving.

This is based on joint work with Bart Bogaerts, Stephan Gocht, Ciaran McCreesh, Magnus O. Myreen, Andy Oertel, and Yong Kiam Tan.

3.18 Aspects of Knowledge for Next Generation Systems

Florian Rabe (Universität Erlangen-Nürnberg, DE)

The Tetrapod model organizes mathematical knowledge into 4+1 aspects, visualized as the corners and the center of a tetrahedral shape. The corners represent fundamentally different ways of assigning semantics, each with an ecosystem of highly specialized tools and large libraries:

- Deduction: proofs, especially if formalized and mechanically verified in proof assistants
- Computation: algorithms, especially if executably implemented in programming languages and computer algebra systems
- Tabulation: systematic lists of examples, especially if encoded as concrete objects stored in databases
- Documentation: human-readable narrative explanations, especially if systematically structured and annotated to enable machine processing

A key novelty of the model is to identify as the central aspect the intersection of the above, called Ontology: names, types, definitions, notations, and properties of mathematical objects, i.e., the information that is critical for knowledge exchange between the dedicated software systems for the other aspects.

This talk gives a high-level overview of the model in discussion-starter style and can be seen as a position statement that next generation systems must invent fundamentally new designs to fully utilize the combination of all aspects.

3.19 Proofs in cvc5: New Directions with AletheLF

Andrew Joseph Reynolds (University of Iowa – Iowa City, US)

Satisfiability Modulo Theories (SMT) solvers are a critical component of many formal methods applications, including for software verification and security analysis. Their soundness is of the utmost importance. While SMT solvers are highly complex systems, some modern SMT solvers now are capable of generating externally checkable proofs. This talk gives the current state of proofs in the SMT solver cvc5. We introduce AletheLF, the new standard format for proofs generated by cvc5. AletheLF is a logical framework based on the SMT-LIB version 3.0 language. It combines the benefits of several previous proof efforts, including a clean syntax, extensibility and integration with other proof formats like DRAT via the use of oracles. We present an initial evaluation of AletheLF, showing the viability of performant proof generation and checking for SMT.

3.20 Using Word Similarities to Guide Resolution

Claudia Schon (Hochschule Trier, DE)

Unlike automated reasoning, human reasoning does not adhere to logical rules exclusively. This is also reflected in the observation of Kahneman that the human mind seems to be based on two integrated systems: a System 1 that works quickly and unconsciously, and a System 2 that works slowly and calculates logically. System 1 embodies intuitions and fast reactions to sensory signals, while System 2 represents deliberate thinking and abstract problem solving. It can be seen as a strength humans have that we have these two very different systems which we are able to combine. And in fact these two systems complement each other very nicely. Hence, the combination of statistical procedures and logical reasoning holds promise for automated reasoning. The meaning of words, like they are captured in Word Embeddings constitutes an important source of information for automated reasoning systems. In knowledge bases where predicate and function symbols align closely with words, these Word Embeddings can be employed. In previous studies, we have demonstrated the successful integration of word similarities into the selection process, where relevant knowledge for a specific query needs to be extracted from a large knowledge base. Additionally, we incorporated Word Embeddings into the selection of the given clause in the given clause algorithm within resolution provers. Initial experimental results indicate that integrating word similarities leads to provers deriving fewer resolvents and maintaining a more focused approach to the query context.

3.21 Proofs for Quantified Boolean Formulas

Martina Seidl (Johannes Kepler Universität Linz, AT)

Quantified Boolean Formulas (QBFs) extend propositional logic by quantifiers over the Boolean variables. As a consequence of having quantifiers, the decision problem of QBF is PSPACE-complete. There is a symmetry between models of true QBFs and counter-models of false QBFs. Both can be represented as binary trees or as sets of Boolean functions, encoding the solutions of application problems that have been translated to QBFs. In practice, those solutions are often extracted from proofs as produced by the QBF solvers.

The landscape of QBF solving paradigms rather heterogeneous, resulting in solvers are based on various proof systems of different strength. In this talk, we review three different proof systems on which recent solvers are built. In particular, we consider Q-resolution for true and false formulas as found in QCDCL, forall-Exp Res as implemented in expansion-based systems as well as QRAT that was developed for recent pre- and inprocessing techniques.

3.22 More than unit equality

Nick Smallbone (Chalmers University of Technology – Göteborg, SE)

Joint work of Nicholas Smallbone, Koen Claessen
Main reference Nicholas Smallbone: "Twee: An Equational Theorem Prover", in Proc. of the Automated Deduction – CADE 28 – 28th International Conference on Automated Deduction, Virtual Event, July 12-15, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12699, pp. 602–613, Springer, 2021.
URL https://doi.org//10.1007/978-3-030-79876-5_35

Equational theorem provers based on Knuth-Bendix completion can solve difficult reasoning problems in, for example, algebra. But the expressive power is limited by the lack of logical connectives. I show that a completion-based prover can reason about practical problems involving connectives with the help of a SAT solver and efficient encodings. I also argue that completion is a useful setting for studying problems in saturation provers, such as how to reason in a goal-directed manner, an important but under-studied problem.

3.23 On hierarchical reasoning and symbol elimination and applications to parametric verification

Viorica Sofronie-Stokkermans (Universität Koblenz, DE)

Joint work of Viorica Sofronie-Stokkermans, Dennis Peuter and Philipp Marohn
Main reference Dennis Peuter, Viorica Sofronie-Stokkermans: "Symbol Elimination and Applications to Parametric Entailment Problems", in Proc. of the Frontiers of Combining Systems – 13th International Symposium, FroCoS 2021, Birmingham, UK, September 8-10, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12941, pp. 43–62, Springer, 2021.
URL https://doi.org//10.1007/978-3-030-86205-3_3
Main reference Dennis Peuter, Viorica Sofronie-Stokkermans: "On Invariant Synthesis for Parametric Systems", in Proc. of the Automated Deduction – CADE 27 – 27th International Conference on Automated Deduction, Natal, Brazil, August 27-30, 2019, Proceedings, Lecture Notes in Computer Science, Vol. 11716, pp. 385–405, Springer, 2019.
URL https://doi.org//10.1007/978-3-030-29436-6_23

We present past and current work on hierarchical symbol elimination.
We first present a goal-oriented symbol elimination method which, given (i) a base theory $\mathcal{T}_0$ allowing quantifier elimination, (ii) an extension $\mathcal{T}_1$ of $\mathcal{T}_0$ with additional function symbols whose properties are axiomatised by a set $\mathcal{K}$ of clauses, (iii) a subset of the additional functions which are considered to be parameters, and (iv) a set G of ground clauses, such that $\mathcal{T}_1 \wedge G$ is satisfiable, computes a universal formula Γ containing symbols in the base theory $\mathcal{T}_0$ and parameters such that $\mathcal{T}_1 \wedge \Gamma \wedge G$ is unsatisfiable. The computation of Γ is done in a hierarchical way, and relies on methods for quantifier elimination in $\mathcal{T}_0$. We identify situations under which the formula Γ computed with our method is the weakest universal formula with the property above, and explain how we used this method for the verification of parametric systems:

1. for generating (weakest) constraints on parameters under which certain properties are guaranteed to be inductive invariants,
2. for iteratively strengthening properties to obtain inductive invariants.

We then briefly present a method for general symbol elimination which uses a constraint resolution calculus obtained from specializing the hierarchical superposition calculus, and explain how we used it – together with goal-oriented symbol elimination – in problems from wireless research theory.

3.24 On Finding Short Proofs

Alexander Steen (Universität Greifswald, DE)

Joint work of Christoph Benzmüller, David Fuenmayor, Alexander Steen, Geoff Sutcliffe
Main reference Christoph Benzmüller, David Fuenmayor, Alexander Steen, Geoff Sutcliffe: "Who Finds the Short Proof?", Logic Journal of the IGPL, p. jzac082, 2023.
URL https://doi.org//10.1093/jigpal/jzac082

The talk reports on an exploration of Boolos' Curious Inference, using higher-order automated theorem provers (ATPs). Surprisingly, only suitable shorthand notations had to be provided by hand for ATPs to find a short proof. The higher-order lemmas required for constructing a short proof are automatically discovered by the ATPs. Given the observations and suggestions in this paper, full proof automation of Boolos' and related examples now seems to be within reach of higher-order ATPs. Preliminary work on automating the synthesis of such shorthand notations is briefly presented.

The talk is based on joint work with Chris Benzmüller, David Fuenmayor and Geoff Sutcliffe [1].

References

1 Christoph Benzmüller, David Fuenmayor, Alexander Steen, Geoff Sutcliffe. *Who Finds the Short Proof? An Exploration of Variants of Boolos' Curious Inference using Higher-order Automated Theorem Provers*. Logic Journal of the IGPL, 2023. DOI: http://doi.org/10.1093/jigpal/jzac082

3.25 TPTP World Standards and Tools for Tarskian and Kripke Interpretations

Geoff Sutcliffe (University of Miami, US), Pascal Fontaine (University of Liège, BE), Jack McKeown (University of Miami, US), and Alexander Steen (Universität Greifswald, DE)

URL https://www.tptp.org/Seminars/TPTPInterpretations/

This talk describes the (new) TPTP World format for representing Tarskian and Kripke interpretations of formulae in classical (FOF, TFF, TXF, THF) and non-classical (NXF, NHF) logics. A technique and implemented tool for verifying models, and a tool for visualizing Tarskian interpretations, are presented. This work provides TPTP World standards that allow interpretations to be shared between components of complex compositional reasoning systems.

3.26 Mechanizing the Splitting Framework

Sophie Tourret (INRIA Nancy – Grand Est, FR)

Joint work of Ghilain Bergeron, Sophie Tourret
Main reference Gabriel Ebner, Jasmin Blanchette, Sophie Tourret: "Unifying Splitting", J. Autom. Reason., Vol. 67(2), p. 16, 2023.
URL https://doi.org//10.1007/S10817-023-09660-8

In this talk, I present the current state of the Isabelle/HOL mechanization efforts by Ghilain Bergeron and myself of the splitting framework by Gabriel Ebner, Jasmin Blanchette and myself. These results include the splitting calculus from section 3 of the framework as well as a partial instance of splitting without backtracking over resolution in FOL. There is still one assumption of this instantiation that is not discharged: the compactness of FOL. Surprisingly, we were unable to find this folklore result in Isabelle/HOL already. I also present the mechanization of this result in Isabelle/HOL via Los's theorem and explain why it is not (yet) usable to discharge the desired assumption of the splitting instance. Finally, I discuss other leads to reach this desired result.

3.27 On the (In-)Completeness of Destructive Equality Resolution in the Superposition Calculus

Uwe Waldmann (MPI für Informatik – Saarbrücken, DE)

Bachmair's and Ganzinger's abstract redundancy concept for the Superposition Calculus justifies almost all operations that are used in superposition provers to delete or simplify clauses, and thus to keep the clause set manageable. Typical examples are tautology deletion, subsumption deletion, and demodulation, and with a more refined definition of redundancy joinability and connectedness can be covered as well. The notable exception is destructive equality resolution, that is, the replacement of a clause $x \not\approx t \vee C$ with $x \notin \mathit{vars}(t)$ by $C\{x \mapsto t\}$. This operation is implemented in state-of-the-art provers, and it is useful in practice, but little is known about how it affects refutational completeness. We demonstrate on the one hand that the naive addition of destructive equality resolution to the standard abstract redundancy concept renders the calculus refutationally incomplete. On the other hand, we present several restricted variants of the superposition calculus that are refutationally complete even with destructive equality resolution.

3.28 The SCL Calculus and its Implementation

Christoph Weidenbach (MPI für Informatik – Saarbrücken, DE)

Main reference Martin Bromberger, Simon Schwarz, Christoph Weidenbach: "SCL(FOL) Revisited", CoRR, Vol. abs/2302.05954, 2023.
URL https://doi.org//10.48550/ARXIV.2302.05954

The talk includes an introduction to the SCL calculus, in particular its version for first-order logic. In addition, I discuss implementation aspects, in particular lifting the CDCL 2-Watched Literal Scheme to first-order logic.

Participants

- Franz Baader
TU Dresden, DE
- Haniel Barbosa
Federal University of Minas Gerais-Belo Horizonte, BR
- Maria Paola Bonacina
University of Verona, IT
- David Déharbe
CLEARSY – Aix-en-Provence, FR
- Martin Desharnais
Max-Planck-Institut für Informatik Saarbrücken, DE
- Clare Dixon
University of Manchester, GB
- Catherine Dubois
ENSIIE – Evry, FR
- Amy Felty
University of Ottawa, CA
- Pascal Fontaine
University of Liège, BE
- Silvio Ghilardi
University of Milan, IT
- Antti Hyvärinen
Certora – Pregassona, CH
- Fuqi Jia
Chinese Academy of Sciences – Beijing, CN
- Chantal Keller
ENS – Gif-sur-Yvette, FR
- Cynthia Kop
Radboud University Nijmegen, NL
- Konstantin Korovin
University of Manchester, GB
- Hanna Lachnitt
Stanford University, US
- Feifei Ma
Chinese Academy of Sciences – Beijing, CN
- Jasper Nalbach
RWTH Aachen, DE
- Claudia Nalon
University of Brasília, BR
- Jakob Nordström
University of Copenhagen, DK & Lund University, SE
- Florian Rabe
Universität Erlangen-Nürnberg, DE
- Andrew Joseph Reynolds
University of Iowa – Iowa City, US
- Claudia Schon
Hochschule Trier, DE
- Stephan Schulz
Duale Hochschule Baden-Württemberg – Stuttgart, DE
- Martina Seidl
Johannes Kepler Universität Linz, AT
- Nick Smallbone
Chalmers University of Technology – Göteborg, SE
- Viorica Sofronie-Stokkermans
Universität Koblenz, DE
- Alexander Steen
Universität Greifswald, DE
- Geoff Sutcliffe
University of Miami, US
- Sophie Tourret
INRIA Nancy – Grand Est, FR
- Uwe Waldmann
MPI für Informatik – Saarbrücken, DE
- Christoph Weidenbach
MPI für Informatik – Saarbrücken, DE
- Akihisa Yamada
AIST – Tokyo, JP

Report from Dagstuhl Seminar 23481

MAD: Microarchitectural Attacks and Defenses

Christopher W. Fletcher[*1], **Marco Guarnieri**[*2], **David Kohlbrenner**[*3], **and Clémentine Maurice**[*4]

1 **University of Illinois – Urbana-Champaign, US.** cwfletch@illinois.edu
2 **IMDEA Software Institute – Madrid, ES.** marco.guarnieri@imdea.org
3 **University of Washington – Seattle, US.** dkohlbre@cs.washington.edu
4 **CNRS – CRIStAL, Lille, FR.** clementine.maurice@inria.fr

Abstract

Microarchitectural attacks subvert the security assumptions many software-level security mechanisms rely upon, thereby threatening the security of our IT systems. These attacks exploit the side-effects (like subtle timing differences in a program's execution time) resulting from a processor's internal optimizations to leak sensitive information and compromise a system's security. Building systems that are resistant against such attacks requires fundamentally rethinking the design of hardware and software security mechanisms.

This seminar gathered together leading researchers that are working on security at the hardware-software interface spanning four different communities: computer security, computer architectures, programming languages and verification, and applied cryptography. The goals were to (1) present a comprehensive overview of current advances in microarchitectural attacks and defenses, (2) foster interaction and future collaboration between researchers from different research communities, and (3) identify interesting research directions and open challenges that need to be addressed to build the next generation of systems that are resistant to microarchitectural attacks.

Seminar November 26 – December 1, 2023 – https://www.dagstuhl.de/23481
2012 ACM Subject Classification Security and privacy → Formal security models; Security and privacy → Security in hardware; Security and privacy → Systems security
Keywords and phrases hardware-software co-design for security, microarchitectural attacks, security architectures, side-channel analysis
Digital Object Identifier 10.4230/DagRep.13.11.151

1 Executive Summary

Christopher W. Fletcher (University of Illinois – Urbana-Champaign, US)
Marco Guarnieri (IMDEA Software Institute – Madrid, ES)
David Kohlbrenner (University of Washington – Seattle, US)
Clémentine Maurice (CNRS CRIStAL – Lille, FR)

Our society relies on a multitude of information systems that generate, process, and store a massive amount of potentially sensitive data. Protecting and regulating the access to this growing collection of data is critical to prevent security breaches and data misuse. For this, information systems deploy many security mechanisms at different levels: from application-level security checks to, for instance, security mechanisms directly implemented in operating systems. These mechanisms are implemented in a layered fashion where mechanisms at a higher level (say, an application-level security check) rely on the security guarantees provided

[*] Editor / Organizer

by lower levels (say, process isolation provided by the operating system). Since the majority of these security mechanisms are implemented in software, their security relies on specific assumptions about how processors execute software.

However, *microarchitectural attacks* have shown, time and again, that many software mechanisms rely on incorrect assumptions about how programs are executed by processors. These attacks, which target the hardware-software interface, exploit the side-effects (like subtle timing differences in a program's execution time) resulting from a processor's internal optimizations to compromise a system's security. Even worse, these attacks clearly highlight that we lack a precise hardware-software interface for security, which is a prerequisite for building trustworthy and reliable security mechanisms.

Scope

The **Dagstuhl Seminar 23481** focused on the topic of **Microarchitectural Attacks and Defenses** (**MAD** for short), a rapidly growing research area focused on discovering, mitigating, and preventing microarchitectural attacks. As an indication of this rapid growth, the Spectre [1] and Meltdown [2] papers – two seminal works (published in 2018) illustrating how microarchitectural attacks can bypass and circumvent many software-level security mechanisms – have jointly attracted more than 4500 citations. Since then, researchers from multiple communities – computer security, computer architectures, programming languages and verification, and applied cryptography – have been working on tackling the challenges posed by microarchitectural attacks. In particular, the MAD community has, so far, been broadly focusing on the following research topics:

Attacks: In terms of attack-oriented research, the MAD community has been focusing on characterizing the microarchitectural side-effects arising in modern processors and on identifying new microarchitectural attacks. In particular, the discovery of new microarchitectural details is often the first step towards developing new attacks. Even though the majority of this research still heavily relies on manual analysis and reverse engineering, researchers started to focus also on the development of approaches and tools to automate the discovery of leaks and attacks.

Hardware and software defenses: The MAD community has also been focusing on the development of defenses and mitigations – spanning the entire spectrum from hardware to software – against microarchitectural attacks. For instance, the community has proposed different ways of modifying current microarchitectures to directly prevent microarchitectural leaks, e.g., by identifying (and delaying) those operations that might result in leaks of sensitive information. In terms of software defenses, instead, the community has been focusing on techniques for securely executing computations even on top of current "leaky" processors, e.g., by relying on compiler-based mitigations to prevent leaks.

Foundations and verification: In terms of foundations and verification, the MAD community has been focusing on three core challenges. First, identifying and formalizing new security abstractions capturing microarchitectural leaks. Second, developing automated techniques for reasoning about microarchitectural leaks in software given high-level leakage models. Third, developing verification techniques for proving the security of processors at register-transfer level against microarchitectural attacks.

Goals

The main goal of the **Dagstuhl Seminar 23481 – MAD: Microarchitectural Attacks and Defenses** was to bring together researchers that work on different, but related, research topics such as

1. microarchitectural and side-channel attacks,
2. software security,
3. computer architectures and hardware security,
4. program verification and formal methods for security, and
5. applied cryptography.

For this, the seminar focused on:

1. Providing an overview of the latest research results related with security at the hardware-software interface with a focus on microarchitectural attacks and defenses.
2. Strengthening the interaction between researchers from different community working on topics relevant to microarchitectural attacks and defenses.
3. Discussing relevant open problems about microarchitectural attacks and defenses, identifying novel insights that can arise by combining results from different research areas, and fostering the collaboration between researchers.

Attendees and seminar's structure

The seminar was attended by 35 researchers with diverse background, spanning all research communities related to MAD: computer security, applied cryptography, computer architectures, and programming languages and verification. The attendees were also a good mix between academia (28 attendees) and industry (7 attendees). This mixture of diverse backgrounds, which was particularly appreciated by many participants, led to many interesting discussions fueled by a wide variety of points of views.

The seminar lasted 4.5 days and it was organized as follows. The first two days were dedicated to establishing a common background for all attendees. This was achieved through overview talks on core MAD topics: (a) microarchitectural attacks and defenses, (b) formal methods and verification, (c) defenses at software and hardware level, and (d) a special session dedicated to Rowhammer attacks and defenses. Each overview topic was covered in 2 talks given by leading researchers on the respective topics. The remaining days were dedicated to contributed talks by the attendees (in the mornings) and small discussion groups (in the afternoons). The discussion groups started from topics proposed by the organizers such as "What are the current capabilities of formal methods approaches and which are the challenges for tackling microarchitectural attacks?", "What is a good methodology for evaluating the security guarantees of microarchitectural defenses?", or "Which interesting future systems/technologies might have implications for microarchitectural security?". On the other days, the discussion was directly driven by the attendees, sometimes continuing on the above topics and sometimes exploring other research questions (e.g., identifying a new taxonomy of microarchitectural attacks).

Future plans

Microarchitectural attacks are here to stay: addressing them requires to fundamentally rethink the design of hardware and software security mechanisms. We believe that the core topics of the MAD Dagstuhl Seminar will be relevant and at the edge of research for a long time. Moreover, the seminar attracted a lot of interest and received positive feedback from the attendees, which particularly appreciated being in contact with leading researchers from other areas working on MAD as well as the presence of both industrial and academic attendees. For these reasons, we believe that this Dagstuhl Seminar should be repeated in the future. Potential improvements for the future editions could be (1) inviting more computer architects and increasing the amount of attendees from industry (in particular, from chip vendors), and (2) dedicating part of the seminar to deep-dives on specific topics.

References

1 Paul Kocher, Jann Horn, Anders Fogh, Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, Michael Schwarz, and Yuval Yarom. Spectre Attacks: Exploiting Speculative Execution. In *Proceedings of the 40th IEEE Symposium on Security and Privacy (S&P 2019)*.

2 Moritz Lipp, Michael Schwarz, Daniel Gruss, Thomas Prescher, Werner Haas, Anders Fogh, Jann Horn, Stefan Mangard, Paul Kocher, Daniel Genkin, Yuval Yarom, and Mike Hamburg. Meltdown: Reading Kernel Memory from User Space. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security 2018)*

2 Table of Contents

3 Overview of Talks

3.1 Microarchitectural defenses in software

Sunjay Cauligi (MPI-SP – Bochum, DE)

In which I discuss various software-based defenses against Spectre attacks. A successful Spectre exploit is comprised of several distinct phases; different mitigations target these different phases, to varying degrees of completeness and performance. In particular, I highlight the Ultimate SLH [1] and Serberus [2] mitigations and how they are able to overcome the subtleties of transient execution.

References

1 Zhang and Barthe and Chuengsatiansup and Schwabe and Yarom. *Ultimate SLH.* USENIX, 2023.
2 Mosier and Nemati and Mitchell and Trippel. *Serberus.* Oakland, 2024.

3.2 Ciphertext Side Channels and their Mitigation

Thomas Eisenbarth (Universität Lübeck, DE)

Joint work of Jan Wichelmann, Anna Pätschke, Luca Wilke, and Thomas Eisenbarth
Main reference Jan Wichelmann, Anna Pätschke, Luca Wilke, Thomas Eisenbarth: "Cipherfix: Mitigating Ciphertext Side-Channel Attacks in Software", in Proc. of the 32nd USENIX Security Symposium (USENIX Security 23), pp. 6789–6806, USENIX Association, 2023.
URL https://www.usenix.org/conference/usenixsecurity23/presentation/wichelmann

In this short talk we discussed memory protection in modern Trusted Execution Environments and the role of logic isolation and/or cryptographic isolation in memory protection. The usage of deterministic encryption enables cipertext side-channel attacks that can be used to extract secrets from constant-time code. Cipherfix patches binaries by masking all writes of secret values with fresh pseudorandom masks, thereby preventing ciphertext side channels in the protected binary. The induced performance overhead is 2x and more for many workloads. It may serve as a lower bound of the expected costs of moving masking-style countermeasures to arbitrary binaries when trying to prevent arbitrary value leakage on server-grade CPUs, as recently exploited by the Hertzbleed attack.

3.3 How can we improve analysis and software mitigation of data-at-rest and value-dependent leakages?

Michael Flanders (University of Washington – Seattle, US)

Joint work of Michael Flanders, Reshabh Sharma, Alexandra Michael, Dan Grossman, David Kohlbrenner
Main reference Michael Flanders, Reshabh Sharma, Alexandra Michael, Dan Grossman, David Kohlbrenner: "Avoiding Instruction-Centric Microarchitectural Timing Channels Via Binary-Code Transformations". ASPLOS 2024, to appear
URL https://homes.cs.washington.edu/ dkohlbre/papers/cio-asplos24.pdf

A group of us at UW have been working on detecting and mitigating data-at-rest and value-dependent leakages caused by novel microarchitectural optimizations. These optimizations include things like simplifiable and bypassable computations, silent stores, the Apple data-memory dependent prefetcher, and others as described in the recent "Opening Pandora's Box ..." paper.

In this talk, I plan to rant about some of the difficulties we faced in implementing leakage analyzers and mitigations in low-level compiler passes and in stand-alone binary analysis tools. These difficulties range from frustrations to soundness issues and arise from improper interfaces and abstractions as well as default assumptions binary analysis tools make that are improper for side-channel analysis. I will briefly discuss our thoughts on solutions but largely want to solicit discussion on better handling of these issues as we see more of these optimizations and accompanying defensive work.

3.4 Attacks from Software Leveraging Microarchitectural Features

Daniel Gruss (TU Graz, AT)

In this talk, we discuss aspects of attacks from software leveraging microarchitectural features decomposed into multiple parts: We discuss the concept of attacks from software and argue that it ranges from attacks with physical access to attacks where the attacker does not even control a single line of code on the victim system. Thus, threat models for attacks from software vary widely. We discuss that the term microarchitecture also is used in different ways in different contexts: Often it refers specifically to the processor microarchitecture but it is increasingly used as a terminological counterpart to architecture, i.e., microarchitecture as the implementation of an architecture, including anything beneath the architectural interface. Finally, we discuss microarchitectural features that facilitate such attacks and the development trends underlying to the scientific progress in this field.

3.5 Software Defenses: What is the correct interface for a hardware "configuration bit"?

David Kohlbrenner (University of Washington – Seattle, US)

With the explosion of novel hardware optimizations, then used for attacks, has come a variety of hardware configuration options for those optimizations. A common approach is a simple on/off bit that can be set in a model specific register (MSR.)

Unfortunately, the preconditions for setting these bits, their effects, and their persistence are decidedly non-uniform. For software-based defenses that intend to use these bits to protect sensitive computation this presents several common problems. Rather than attempt to solve each configuration case on its own, we ask what an ideal simple configuration interface would look like for a compiler-based hardening scheme.

3.6 What can speculative execution learn from exploitation?

Anil Kurmus (IBM Research-Zurich, CH)

We draw parallels between speculative execution attacks and memory errors. Exploitation of memory errors has a long history, starting from the 1972 Anderson report. While the problem is much more close to being solved in a principled and practical way 50 years later, we have not quite succeeded. What are the lessons we can learn and apply for speculative execution defenses? A few topics of further discussion include "minimum viable patching" vs. principled defenses, attack chaining, attack reliability and portability, taxonomies inspired by memory errors.

3.7 Modeling and Detecting Microarchitectural Leaks

Boris Köpf (Microsoft Research – Cambridge, GB)

Speculative execution attacks such as Spectre and Meltdown exploit microarchitectural optimizations to leak information across security domains. These vulnerabilities often stay undetected for years, because we lack the tools for systematic analysis of CPUs to find them.

In this talk I presented leakage contracts as a way to specify speculative leaks together with Revizor, a tool that can automatically test CPUs against these specifications. I gave examples of how this approach can be used to detect large classes of known and unknown leaks in recent x86 CPUs.

3.8 RowHammer, RowPress and Beyond: Can We Be Free of Bitflips (Soon)?

Onur Mutlu (ETH Zürich, CH)

Main reference Onur Mutlu, Ataberk Olgun, A. Giray Yağlıkcı: "Fundamentally Understanding and Solving RowHammer", in Proc. of the 28th Asia and South Pacific Design Automation Conference, ASPDAC '23, ACM, 2023.
URL http://dx.doi.org/10.1145/3566097.3568350

We will examine the RowHammer problem in Dynamic Random Access Memory (DRAM), the first example of how a circuit-level failure mechanism can cause a practical and widespread system security vulnerability. RowHammer is the phenomenon that repeatedly accessing a row in a modern DRAM chip predictably causes bitflips in physically-adjacent rows. Building on our initial fundamental work that appeared at ISCA 2014, Google Project Zero demonstrated that this hardware phenomenon can be exploited by user-level programs to gain kernel privileges. Many other works demonstrated other attacks exploiting RowHammer, including remote takeover of a server vulnerable to RowHammer, takeover of a mobile device by a malicious user-level application, and destruction of predictive capabilities of commonly-used deep neural networks.

Unfortunately, the RowHammer problem still plagues cutting-edge DRAM chips, DDR4 and beyond. Based on our recent characterization studies of more than 1500 DRAM chips from six technology generations that appeared at ISCA 2020 and MICRO 2021, we show that RowHammer at the circuit level is getting much worse, newer DRAM chips are much more vulnerable to RowHammer than older ones, and existing mitigation techniques do not work well. We also show that existing proprietary mitigation techniques employed in DDR4 DRAM chips, which are advertised to be Rowhammer-free, can be bypassed via many-sided hammering (also known as TRRespass & Uncovering TRR).

In this talk, we will provide an overview of RowHammer research in academia and industry, with a special focus on recent works that rigorously analyze real chip characteristics and introduce promising solution ideas. We will discuss the effect of RowHammer on High-Bandwidth Memory (HBM) chips and introduce and analyze RowPress, which is a fundamentally different read disturbance phenomenon that also affects all DRAM chips. RowPress greatly (e.g., by 100X) reduces the activation count required to induce bitflips, by keeping an activated row open for a long time. We will also discuss what other problems may be lurking in DRAM and other types of memory, which can potentially threaten the foundations of reliable and secure systems, as memory technologies scale to higher densities. We will conclude by describing and advocating a principled approach to memory robustness (including reliability, security, safety) research that can enable us to better anticipate and prevent such vulnerabilities.

- A short accompanying paper, which appeared at ASP-DAC 2023, can be found here and serves as recommended reading: "Fundamentally Understanding and Solving Row-Hammer" `https://arxiv.org/abs/2211.07613`.
- Slides: `https://people.inf.ethz.ch/omutlu/pub/onur-DagStuhl-MAD-RowHammer-28-November-2023.pdf`
- A similar talk online on Youtube: `https://www.youtube.com/watch?v=0W7YRRhnunw`

3.9 Security of PIM (Processing-in-Memory) Systems

Onur Mutlu (ETH Zürich, CH)

Main reference Onur Mutlu, Saugata Ghose, Juan Gómez-Luna, Rachata Ausavarungnirun: "A Modern Primer on Processing in Memory", CoRR, Vol. abs/2012.03112, 2022.
URL https://arxiv.org/abs/2012.03112

PIM systems, which enable various types of computation near (or using) memory structures, are gaining traction. We posit that, on the one hand, different types of PIM systems can cause new security issues, exacerbate known issues, or cause new complications related to security. On the other hand, PIM systems can be used to improve security properties by exposing data less, performing security critical functions in memory, or defining new (and physically smaller) trust boundaries in the system. This talk discusses challenges and opportunities in security of PIM systems.

Some related resources are mentioned below:

- A 2-page overview paper from DAC 2023: "Memory-Centric Computing", `https://arxiv.org/abs/2305.20000`
- A short vision paper from DATE 2021: "Intelligent Architectures for Intelligent Computing Systems", `https://arxiv.org/abs/2012.12381`
- A longer survey of modern memory-centric computing ideas and systems (updated August 2022): "A Modern Primer on Processing in Memory", `https://arxiv.org/abs/2012.03112`
- Slides: `https://people.inf.ethz.ch/omutlu/pub/onur-Dagstuhl-PIM-Security-28-November-2023.pdf`

3.10 Practical Rowhammer Attacks and Defenses

Kaveh Razavi (ETH Zürich, CH)

This lecture covers the reverse engineering of in-DRAM Target Row Refresh mechanisms and uses the insights in the development of advanced Rowhammer attacks that bypass these mitigations and the development of principled and scalable alternatives that are secure against these attacks.

3.11 The Gates of Time: Improving Cache Attacks with Transient Execution

Eyal Ronen (Tel Aviv University, IL)

Joint work of Daniel Katzman, William Kosasih, Chitchanok Chuengsatiansup, Eyal Ronen, Yuval Yarom
Main reference Daniel Katzman, William Kosasih, Chitchanok Chuengsatiansup, Eyal Ronen, Yuval Yarom: "The Gates of Time: Improving Cache Attacks with Transient Execution", in Proc. of the 32nd USENIX Security Symposium (USENIX Security 23), pp. 1955–1972, USENIX Association, 2023.
URL https://www.usenix.org/conference/usenixsecurity23/presentation/katzman

For over two decades, cache attacks have been shown to pose a significant risk to the security of computer systems. In particular, a large number of works show that cache attacks provide a stepping stone for implementing transient-execution attacks. However, much less effort has been expended investigating the reverse direction—how transient execution can be exploited for cache attacks. In this work, we answer this question.

We first show that using transient execution, we can perform arbitrary manipulations of the cache state. Specifically, we design versatile logical gates whose inputs and outputs are the caching state of memory addresses. Our gates are generic enough that we can implement them in WebAssembly. Moreover, the gates work on processors from multiple vendors, including Intel, AMD, Apple, and Samsung. We demonstrate that these gates are Turing complete and allow arbitrary computation on cache states, without exposing the logical values to the architectural state of the program.

We then show two use cases for our gates in cache attacks. The first use case is to amplify the cache state, allowing us to create timing differences of over 100 millisecond between the cases that a specific memory address is cached or not. We show how we can use this capability to build eviction sets in WebAssembly, using only a low-resolution (0.1 millisecond) timer. For the second use case, we present the Prime+Scope attack, a variant of Prime+Probe that decouples the sampling of cache states from the measurement of said state. Prime+Store is the first timing-based cache attack that can sample the cache state at a rate higher than the clock rate. We show how to use Prime+Store to obtain bits from a concurrently executing modular exponentiation, when the only timing signal is at a resolution of 0.1 millisecond.

3.12 Rowhammer: Learnings from Designing Defenses and Outlook For the Future

Gururaj Saileshwar (University of Toronto, CA)

Rowhammer is a vulnerability affecting newer generations of DRAM (DDR3,DDR4,LPDDR4) where rapid activations of DRAM rows causes bit-flips in neighboring rows. Moreover, recent victim focused mitigation (refreshing victims neighboring aggressor rows) implemented in DDR4 have also been defeated by new attacks.

This talk discusses three recent Rowhammer mitigations proposing new aggressor-focused mitigations – Randomized Row Swap (RRS) [1], Scalable & Secure Row Swap (SRS) [2], and AQUA [3]. Based on learnings from these defenses, this talk summarizes the outlook for Rowhammer mitigations going forward.

References

1 Gururaj Saileshwar, Bolin Wang, Moinuddin Qureshi, Prashant J. Nair. *Randomized Row-Swap: Mitigating Row Hammer by Breaking Spatial Correlation between Aggressor and Victim Rows.* ASPLOS 2022. https://dl.acm.org/doi/10.1145/3503222.3507716

2 Jeonghyun Woo, Gururaj Saileshwar, Prashant J. Nair. *Scalable and Secure Row-Swap: Efficient and Safe Row Hammer Mitigation in Memory Systems.* HPCA 2023. https://www.computer.org/csdl/proceedings-article/hpca/2023/10070999/1LMbzYX6Uww

3 Anish Saxena, Gururaj Saileshwar, Prashant J. Nair, Moinuddin Qureshi. *AQUA: Scalable Rowhammer Mitigation by Quarantining Aggressor Rows at Runtime.* IEEE MICRO 2022. https://ieeexplore.ieee.org/document/9923789

3.13 Verified Software Security Down to Gates

Caroline Trippel (Stanford University, US)

Hardware-software (HW-SW) contracts are critical for high-assurance computer systems design and an enabler for software design/analysis tools that find and repair hardware-related bugs in programs. E.g., memory consistency models define what values shared memory loads can return in a parallel program. Emerging security contracts define what program data is susceptible to leakage via hardware side-channels and what speculative control- and data-flow is possible at runtime. However, these contracts and the analyses they support are useless if we cannot guarantee microarchitectural compliance, which is a "grand challenge." Notably, some contracts are still evolving (e.g., security contracts), making hardware compliance a moving target. Even for mature contracts, comprehensively verifying that a complex microarchitecture implements some abstract contract is a time-consuming endeavor involving teams of engineers, which typically requires resorting to incomplete proofs.

Our work takes a radically different approach to the challenge above by synthesizing HW-SW contracts from advanced (i.e., industry-scale/complexity) processor implementations. In this talk, I present our work on: synthesizing security contracts from processor specifications written in Verilog; designing compiler approaches parameterized by these contracts that can find and repair hardware-related vulnerabilities in programs; and updating hardware microarchitectures to support scalable verification and efficient security-hardened programs.

3.14 Interrupt-Driven Attacks and Defenses for Microarchitectural Security

Jo Van Bulck (KU Leuven, BE)

Microarchitectural side-channel attacks often face challenges due to limited temporal resolution. Researchers have innovatively employed timer and inter-processor interrupts to temporarily halt victim programs, allowing precise probing of microarchitectural buffers. This technique, while not exclusive to Trusted Execution Environments (TEEs), has demonstrated particular efficacy in such environments.

In this presentation, I share my experiences developing SGX-Step, an open-source framework enabling precise interrupt capabilities within Intel's SGX TEE. I outline specific attack applications of SGX-Step in recent years and its significant impact on the design of effective defenses. Drawing from a thorough root-cause analysis, I explain our collaboration with Intel to devise a hardware-software co-design effectively countering SGX-Step's ability to single-step a victim enclave. Additionally, I highlight our efforts in designing defenses across the system stack for embedded MSP430 Sancus TEE processors. The talk aims to provide insights into interrupt-driven attack evolution and key design choices for mitigating their effects.

3.15 Hardware attacks and defenses: intro and setting the scene

Ingrid Verbauwhede (KU Leuven, BE) and Jesse De Meulemeester (KU Leuven, BE)

In this presentation, we introduce hardware, i.e. physical attacks on electronic circuits. With physical security, we mean sensitive information that can be obtained by monitoring or disturbing the physical behavior of the electronic circuit. A first class of attacks are based on passive observation of the data-dependent variations in timing, power consumption or EM emanation. The strength of these attacks is that the device under attack is not aware that it is being observed. A second class of attacks, called fault attacks, actively manipulate the behavior of the integrated circuits. Examples are clock or power glitching, cooling or heating, laser or EM injection, row hammering and more. The effect of these attacks could be transient or permanent. In a second part of the presentation, we give an overview of the effort and lab set-up which is needed to perform these attacks, ranging from simple cheap power probes to laser and FIB set-ups, both for passive and active attacks. In the last part we discussed countermeasures to protect against passive side-channel and active fault attacks against crypto implementations. Countermeasures are split into two main classes. One is hiding, where the goal is to reduce the signal-to-noise ratio of sensitive data. Examples are logic styles as WDDL, clock jitter, instruction shuffling, etc. The second is masking, where sensitive data is randomly split in shares. Operations then work on randomized data and the signal traces do not contain sensitive data that can directly be correlated to the sensitive data. Higher order attacks require higher order masking, i.e. split in a larger number of shares. Countermeasures against fault attacks include on-chip sensors at the circuit level, redundance and error correcting codes at the algorithm level. Unfortunately, countermeasures against one class of attacks might make the circuit vulnerable to the other class of attacks. Countermeasures resistant to both classes of attacks remain a big research challenge.

4 Working groups

4.1 Tools for Program Analysis

Billy Brumley (Rochester Institute of Technology, US), Steve Kremer (INRIA Nancy – Grand Est, FR), Moritz Lipp (Amazon Web Services – Wien, AT), Nicky Mouha (NIST – Gaithersburg, US), Alastair Reid (Intel – London, GB), and Jan Reineke (Universität des Saarlandes – Saarbrücken, DE)

A prerequisite to identify microarchitectural attacks and protect against them is to explore various tools that are available to understand program properties. This working group focused on getting some hands-on experience with two specific tools: KLEE and CodeQL.

KLEE is a dynamic symbolic execution engine that can be used to automatically reason about software programs. For example, a programmer can add the `klee_assert(a + b >= a)` statement to determine if there exist values that would cause the addition `a + b` to overflow (thereby making the assertion fail). As an example, KLEE was used to analyze a possible integer overflow in code that was present in OpenSSL's HKDF implementation.

CodeQL is a static analysis tool that can perform SQL-like queries to look for specific patterns in source code. An application of CodeQL was explored to detect the pattern that caused a buffer overflow vulnerability in an earlier version of the "official" SHA-3 implementation.

5 Open problems

5.1 Microarchitectural Side-Channel Mitigations for Serverless Applications

Aastha Mehta (University of British Columbia – Vancouver, CA)

Joint work of Yayu Wang, Aastha Mehta

Most of the prior work has focused on microarchitectural side-channel mitigations for cryptographic applications. While cryptography is an important class of applications, we explore microarchitectural side-channel vulnerabilities in other application domains. Specifically, we develop automatic mitigations for serverless applications hosted in cloud platforms. Serverless platforms rely on resource multiplexing among tenants for economies of scale and therefore, coarse-grained resource-partitioning based mitigations are inefficient. Instead, we investigate constant-time execution technique as a principled solution.

Participants

- Gilles Barthe
MPI-SP – Bochum, DE
- Thomas Bourgeat
EPFL – Lausanne, CH
- Billy Brumley
Rochester Institute of Technology, US
- Sunjay Cauligi
MPI-SP – Bochum, DE
- Chitchanok Chuengsatiansup
The University of Melbourne, AU
- Jesse De Meulemeester
KU Leuven, BE
- Thomas Eisenbarth
Universität Lübeck, DE
- Michael Flanders
University of Washington – Seattle, US
- Christopher W. Fletcher
University of Illinois – Urbana-Champaign, US
- Anders Fogh
Intel – Neubiberg, DE
- Daniel Gruss
TU Graz, AT
- Marco Guarnieri
IMDEA Software Institute – Madrid, ES
- Boris Köpf
Microsoft Research – Cambridge, GB
- David Kohlbrenner
University of Washington – Seattle, US
- Steve Kremer
INRIA Nancy – Grand Est, FR
- Anil Kurmus
IBM Research-Zurich, CH
- Moritz Lipp
Amazon Web Services – Wien, AT
- Aastha Mehta
University of British Columbia – Vancouver, CA
- Nicky Mouha
NIST – Gaithersburg, US
- Onur Mutlu
ETH Zürich, CH
- Hamed Nemati
CISPA – Saarbrücken, DE
- Yossi Oren
Ben Gurion University – Beer Sheva, IL
- Riccardo Paccagnella
Carnegie Mellon University – Pittsburgh, US
- Kaveh Razavi
ETH Zürich, CH
- Alastair Reid
Intel – London, GB
- Jan Reineke
Universität des Saarlandes – Saarbrücken, DE
- Tamara Rezk
INRIA – Sophia Antipolis, FR
- Eyal Ronen
Tel Aviv University, IL
- Gururaj Saileshwar
University of Toronto, CA
- Michael Schwarz
CISPA – Saarbrücken, DE
- Mark Silberstein
Technion – Haifa, IL
- Caroline Trippel
Stanford University, US
- Jo Van Bulck
KU Leuven, BE
- Ingrid Verbauwhede
KU Leuven, BE
- Hugo Vincent
Arm – Cambridge, GB

Report from Dagstuhl Seminar 23482

Social XR: The Future of Communication and Collaboration

Mark Billinghurst*[1], Pablo Cesar*[2], Mar Gonzalez-Franco*[3], Katherine Isbister*[4], Julie Williamson*[5], and Alexandra Kitson†[6]

1 **University of South Australia – Adelaide, AU.** mark.billinghurst@unisa.edu.au
2 **CWI – Amsterdam, NL.** p.s.cesar@cwi.nl
3 **Google – Seattle, US.** margonzalezfranco@gmail.com
4 **University of California at Santa Cruz, US.** kisbiste@ucsc.edu
5 **University of Glasgow, GB.** julie.williamson@glasgow.ac.uk
6 **Simon Fraser University – Surrey, CA.** alexandra_kitson@sfu.ca

Abstract

We are rapidly moving towards a hybrid world where communication and collaboration occur in reality, virtuality, and everywhere in-between. But, are current technologies ready for such a shift? Social Extended Reality (XR) systems promise to overcome the limitations of current real-time teleconferencing systems, enabling a better sense of immersion, enhancing the sense of presence, and fostering more successful interpersonal interactions. The possibility for familiar, meaningful, and strategically heightened social interaction in XR has positioned immersive technology as the future of real-time communication and collaboration. This Dagstuhl Seminar gathered academics and practitioners from different disciplines to address the open challenges of immersive interaction including the ethical, legal and societal aspects of possible futures. Participants shared their work through rapid talks and XR demos. The seminar organizers provided provocation talks before small groups convened to discuss three topics over three days: XR design approaches, ethics and values; capturing and modelling; and proxemics, metrics, instrumentation and evaluation. We conclude with a set of grand challenges in the field of social XR in the areas of empathic computing, blended reality, assets and datasets, and survey instruments.

Seminar November 26 – December 1, 2023 – https://www.dagstuhl.de/23482

2012 ACM Subject Classification Human-centered computing → Mixed / augmented reality; Human-centered computing → Collaborative and social computing

Keywords and phrases Social XR, Augmented Reality, Virtual Reality, Extended Reality, Social Computing

Digital Object Identifier 10.4230/DagRep.13.11.167

* Editor / Organizer
† Editorial Assistant / Collector

1 Executive Summary

Mark Billinghurst (University of South Australia, Adelaide, AU)
Pablo Cesar (CWI, Amsterdam, NL)
Mar Gonzalez-Franco (Google, Seattle, US)
Katherine Isbister (University of Santa Cruz, US)
Alexandra Kitson (Simon Fraser University, Surrey, CA)
Julie Williamson (University of Glasgow, GB)

This Dagstuhl Seminar focused on Social XR and the future of communication and collaboration, with a particular interest on:

- Capturing and modelling of humans, ensuring realistic representation of the users and thus allowing for realistic and immersive experiences;
- Digital proxemics and social metrics, that help and enrich communication and collaboration between the participants;
- Instrumentation and evaluation, focusing on the possibility of evaluating and monitoring the experience of the users;
- Principles of Social XR, for making sure that the right values and principles are followed;
- Exploration of design approaches for Social XR, that support communication and connection by enabling and strategically heightening social signalling and dynamics.

To start the seminar, each participant presented relevant social XR research through rapid talks (see Section 2 and Figure 2), which were then used to finalize the topics to discuss on the remaining days of the seminar. We settled on three topics that participants in small groups would focus on, led by members of the organizing committee with a provocation at the start of each day:

TUESDAY: Social XR Design Approaches, Ethics, and Values led by Katherine Isbister and Alexandra Kitson,
WEDNESDAY: Capturing and Modeling Social XR led by Mark Billinghurst and Mar Gonzalez-Franco,
THURSDAY: Proxemics, Metrics, Instrumentation, and Evaluation of Social XR led by Pablo Cesar and Julie Williamson,
THURSDAY: Grand Challenges of Social XR led by the organizers.

As a major result of the seminar, we identified the following grand challenges:

1. Subjectivity of scientific evaluation of empathy
2. Ethical concerns of sharing physiological data and social XR relationships
3. Ethics of the growing digital divide
4. Blending realities, beyond visual and audio
5. Semantic understanding of the physical and social context
6. Social stitching to create a cohesive scene or world
7. Preserving privacy given the increasing fidelity of capture devices
8. Tension between transparency and social superpowers
9. Devising a shared platform that facilitates collaborative recording, replaying, and immersive experiences
10. New metrics and questionnaires for social XR

In addition to the rapid talks and topic discussions, participants shared demos of their work on Tuesday:

- Alexandra Kitson – Embodied Telepresent Connection: An interactive art piece designed to support connection and pseudohaptics through visuals and audio `http://ispace.iat.sfu.ca/project/etc/`
- Alijosa Smolic – Volograms: record a video and turn it into an AR experience `https://www.volograms.com/`
- Anthony Steed – Ubiq: a free, open-source networking library for research, teaching and development `https://ubiq.online/`
- Zerrin Yumak – FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis using Self-Supervised Speech Representation Learning `https://github.com/galib360/FaceXHuBERT`

Social activities in the music room, cellar, games room, and sauna led to some discussions around capturing and modelling leading into Wednesday's session (see **this social media post for some examples**) as well as an impromptu research study on cross-reality asymmetrical co-located social games by playing two games: **DAVIGO** and **Acron**.

Figure 1 Participants enjoy social events around Dagstuhl.

In terms of outputs and future collaborations, we plan to share our findings in an opinion article or forum. We have analyzed, written, and submitted the results of the impromptu research study in the cellar and games room to a top-tier conference in our field. Additionally, we have discussed a potential book on social XR with the seminar participants based on the topics of this seminar. We plan to hold follow-up events and workshops at relevant conferences to further explore the grand challenges that we identified through this seminar.

2 Table of Contents

Figure 2 Participants present their work on social XR. **Left:** Julie Williamson talks about the seminar themes. **Center:** participants listen to presentations and provocations. **Right:** Anthony Steed presents Ubiq.

3 Overview of Talks

3.1 Shared Realities in Social XR

Sun Joo (Grace) Ahn (University of Georgia – Athens, US; sjahn@uga.edu)

I am the founding director of the Center for Advanced Computer-Human Ecosystems at University of Georgia (https://www.ugavr.com). Our work has looked at how virtual experiences can transfer into the physical world to continue changing attitudes, behaviors, and worldviews. In particular, social XR can provide a common ground of shared experiences to multiple users, leading to stronger group cohesion. In addition to sharing experiences interpersonally, social XR allows human users to share experiences with virtual agents. Although prior literature in XR has generally focused on embodiment to provide users to shared experiences of others, emerging evidence points to shared experiences that allow users to walk alongside others being more effective than briefly embodying an avatar body. Our recent research project investigates how social XR with other users and virtual agents can establish a sense of shared reality and lead to the generation of collective minds and empathy. We will present how we have integrated community-based participatory research approaches in developing a prototype virtual experience of sharing the reality of redlining, a past zoning policy that has created long-term structural inequity, leading to negative public health issues for marginalized communities across the US.

3.2 Towards Volumetric Video Conferencing

Pablo Cesar (CWI – Amsterdam, NL; P.S.Cesar@cwi.nl)

I lead the Distributed and Interactive Systems (DIS) group at Centrum Wiskunde & Informatica, CWI, (The National Research Institute for Mathematics and Computer Science in the Netherlands) and I am Professor ("Human-Centered Multimedia Systems Chair) at TU Delft, in the Multimedia Computing group. The work in the group combines human-computer interaction and multimedia systems, focusing on facilitating and improving the way people use interactive systems and how people communicate with each other. We combine data science with a strong human-centric, empirical approach to understand the experience of users. This enables us to design and develop next generation intelligent and empathic systems. With Social Extended Reality (XR) emerging as a new medium, where users can remotely experience immersive content with others, the vision of a true feeling of 'being there together' has become a realistic goal. Together with my group, we have been working towards such a goal, including the development and deployment of an open-source volumetric video conference system, VR2Gather. The system allows for highly-realistic digital humans, based on point cloud capture, encoding, and transmission. Based on results from practical case studies in different sectors (e.g., cultural heritage, performing arts...) in projects such as 5D Culture, TRANSMIXR, and MediaScape XR, we can better understand the existing challenges and to discover the opportunities of this new medium.

3.3 Ubiquitous Metadata: Integrated Fingerprints for Real-World Object Identification and Augmentation

M. Doga Dogan (MIT – Cambridge, US; doga@mit.edu)

In the evolving landscape of immersive experiences, my research focuses on seamlessly integrating physical objects with their digital counterparts through innovative identification, sensing, and tagging methods. By embedding machine-readable tags that convey an object's identity, origin, and function, I establish gateways to "ubiquitous metadata" in the real world. This concept, akin to digital file metadata, empowers users to augment real-world objects with multimedia content, foster interactive experiences in AR/VR, and retrieve contextual information via digital product passports.

My work intersects with Social XR by enabling collaborative interactions in real-world scenarios. By contextualizing and identifying objects in XR, users may for example collaboratively annotate, share, and interact with their physical surroundings. Whether taking or checking notes during shopping, or enhancing home communication with dynamic, interactive messages, my research aims to enhance the intersection of AR and real-world collaboration. During the seminar, I am eager to explore diverse Social XR applications and address critical considerations such as privacy implications in this evolving landscape.

3.4 From Multi-modal to Multi-device interactions in XR

Eric J. Gonzalez (Google – Seattle, US; ejgonz@google.com)

I am a researcher in the Blended Interactions Research & Devices Lab at Google, where I lead the exploration of mutli-modal and multi-device experiences for XR. Currently, my work focuses on how we can leverage existing ecosystems of devices (e.g., smartphones, smartwatches) to augment and supplement natural input techniques (e.g., gaze, gesture, touch). My work connects to Social XR by enabling collaborative interaction scenarios mediated by familiar devices and modalities. For example: in the near term, it is very likely that users in immersive XR (and those around them) will have a smartphone in their pocket.

Supporting device-mediated interactions in XR not only allows immersed users to leverage the sensing and computation offered by their phone (e.g., for precise multi-touch input), but it also enables surrounding collaborators to view and interact with shared XR content through their own devices. I am excited to discuss the future of input in XR as well as other interesting topics such as AI-mediated collaborative experiences.

3.5 Perceptual Manipulations in XR During Face-to-Face Social Interactions

Jan Gugenheimer (TU Darmstadt, DE & Telecom Paris, FR; jan.gugenheimer@TU Darmstadt.de)

My research focuses on two directions at the intersection of XR and HCI: 1: Understand and provide software and hardware solutions on how XR technology has to change to be integrated into an everyday usage scenario (ubiquitous XR) and 2: Understand what potential negative and abusive scenarios (perceptual manipulations, dark patterns) in this future of ubiquitous XR could arise and how we can start shaping the technology to avoid those. In the field of social XR we started to explore how AR technology could impact face-to-face social interactions during ubiquitous XR usage. (https://dl.acm.org/doi/abs/10.1145/3411764.3445597, https://dl.acm.org/doi/abs/10.1145/3491102.3502140). I think one of the core challenges in XR is to understand and leverage its ability to impact the user's perception of themselves and the real world. The biggest difference between traditional digital media (smartphones and PCs) in contrast to XR is its ability to alter the user's perception of the real world (not only the digital). This comes with so many exciting possibilities to improve the technology (e.g., redirected walking, haptic illusions) but also with potential risks (perceptual manipulations). The methods we use in our research are partially grounded in traditional engineering approaches (prototyping and empirical evaluations) but are now extended more and more with design research methods (speculative design, design fiction). I am very eager to discuss how the abilities of XR to alter the user's perception of the world can impact (positive and negative) face-to-face social interactions in the future.

3.6 Meaningful Social VR Environments

Linda Hirsch (LMU Munich, DE; linda.hirsch@ifi.lmu.de)

My research focuses on implicitly increasing socio-cultural connectedness and awareness in shared environments. This includes tracking, moderating, and visualizing users' activities in VR over different periods (e.g., what happened the day before or two years ago). By this, meaningful user experiences are created by contextualizing VR interactivity and fostering a deeper connection with the VR environment and other users. The means to realize a deeper connection are endless. Yet, the challenge is based on balancing the amount and quality of information, the communication channels, and the translation of comprehensible information. For this, I apply methods and theory from anthropology, materials experience design, and environmental psychology research, e.g., meaning of place framework of place attachment, in addition to common HCI methods. Choosing the "right" method is context-dependent, e.g., are we looking at physical or virtual reality contexts? In addition, it is very important to consider the long-term effects and the "history" of a shared VR space regarding its socio-cultural effects on the virtual, physical, and mixed social reality.

3.7 Designing Social VR Meeting Spaces

Katherine Isbister (University of California, Santa Cruz, US; katherine.isbister@ucsc.edu)

For the last several years my research team has been building Research through Design prototypes of social VR meeting spaces, taking a "beyond being there" approach, with funding first from Mozilla, then from the US National Science Foundation. We've written papers about the general approach (see `https://dl.acm.org/doi/10.1145/3411763.3450377` and `https://www.tandfonline.com/doi/abs/10.1080/07370024.2021.1994860`) and have released a toolkit that others are welcome to use (`http://info.socialsuperpowers.net/`) from this work. More recently, we've received a grant from the Sloan Foundation to build social VR prototypes to support scientific sensemaking around spatialized data.

3.8 Social Communication and Connection in XR

Alexandra Kitson (Simon Fraser University – Surrey, CA; akitson@sfu.ca)

Joint work of Alexandra Kitson, John Desnoyers-Stewart, Ekaterina R. Stepanova, Pinyao Liu, Patrick Parra Pennefather, Vladislav Ryzhov, Bernhard E. Riecke, Alissa Antle, Petr Slovak, Katherine Isbister, Ashu Adhikari, Kenneth Karthik

Main reference John Desnoyers-Stewart, Ekaterina R. Stepanova, Pinyao Liu, Alexandra Kitson, Patrick Parra Pennefather, Vladislav Ryzhov, Bernhard E. Riecke: "Embodied Telepresent Connection (ETC): Exploring Virtual Social Touch Through Pseudohaptics", in Proc. of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA '23, Association for Computing Machinery, 2023.
URL http://dx.doi.org/10.1145/3544549.3585843

I design, develop, implement, and evaluate VR applications for both social transformation and emotional well-being. Two projects that relate to Social XR: (1) "Embodied Telepresent Connection" gives the illusion of social touch and bodily connection through visuals, sounds, biosignals, and embodied metaphors in VR, connecting distanced people in the same virtual space. (2) Go-along interviews in VRChat with adolescents to better understand the social spaces youth are using and the distinctive features of those spaces that contribute to successful emotion regulation. Some of the core challenges I see in the field:

1. Interaction and communication with others in social XR.
2. Representing people in a virtual space, including across mixed platforms.
3. Safeguards and spaces for vulnerable people (e.g., children) in social XR.
4. Design tools that aren't prototyping.

I will share my experiences using participatory and embodied design methods, as well as ideas around pseudo-haptics and biosignal integration to enhance social communication and connection in XR. I'm most interested in discussing design approaches, values, and ethics.

3.9 Augmented Social Perception

Kai Kunze (Keio University – Yokohama, JP; kai.kunze@pm.me)

My research is centered on the exploration and development of technology tool-sets designed to augment human capabilities and overcome our physical and cognitive limitations. The human head, being the center of our senses, vital signs, and actions, presents an ideal location for simultaneous sensing and interactions of assistance applications. By integrating sensing and interaction modalities into the form factor of eyeglasses, we can create multi-purpose wearable monitoring and assistance devices.

3.10 Designing and Evaluating User Experiences in Social Virtual Reality (VR)

Jie LI (EPAM – Hoofddorp, Noord-Holland, NL; jie_li@epam.com)

My interest in Social VR is focused on designing novel user experiences and developing metrics and methods to understand and measure aspects such as user engagement, cognitive load, enjoyment, quality of interaction, and social connectedness. As a researcher in the industry who also collaborates closely with academic researchers, I often observe a disconnect between the two worlds. Industry projects often prioritize application and market readiness, sometimes neglecting the foundational reasons for designing and developing social VR experiences. In contrast, academia usually concentrates on fundamental research and may overlook the practical application of lab innovations for everyday public use. The future of social XR calls for collaborative efforts between academia and industry, demanding not only advanced fundamental research but also sophisticated user experience design. This could include the development of a standardized design system for XR, containing well-tested design components that can be directly used to create the basic user experience architecture. Such collaboration will ensure that diverse users are engaged and included, with accessible hardware and software, complemented by open-sourced evaluation methods, metrics, or shared platforms to facilitate the consistent improvement of user experiences.

3.11 Philosophy and XR Technology

Neil McDonnell (University of Glasgow, GB; Neil.McDonnell@glasgow.ac.uk)

I am a philosopher at the University of Glasgow. I used to work in the 3D viz industry and as a result, I do a lot of interdisciplinary work and lead major projects concerning XR deployments. I have a practical eye for XR deployment issues in research and education. I write about what causation is, whether virtual things are real or valuable, policy papers about XR and education, and the nature of evidence in safety systems. I approach this wide range of topics with analytic philosophy training from metaphysics. I am not an ethicist, and I do not think at all about the meaning of life. I think the big issues around social XR are about access, acceptance and adoption. Who can access and who can not – who are we leaving behind? Why are so many people so resistant to this incredible technology? We need to answer the first two before widescale ubiquitous adoption will be achieved.

3.12 The Empathic Metaverse: An Assistive Bioresponsive Platform for Emotional Experience Sharing in Social XR

Yun Suen Pai (The University of Auckland, NZ; yun.suen.pai@auckland.ac.nz)

My research explores the social impact of XR and how it can be used to assist, augment and understand others. The Metaverse is poised to be a future platform that redefines what it means to communicate, socialize, and interact with each other. Yet, it is important for us to consider avoiding the pitfalls of social media platforms we use today; cyberbullying, lack of transparency and an overall false mental model of society. In this seminar, I would like to discuss about the Empathic Metaverse, a virtual platform that prioritizes emotional sharing for assistance. It aims to cultivate prosocial behaviour, either egoistically or altruistically, so that our future society can better feel for each other and assist one another. To achieve this, I propose the platform to be bioresponsive; it reacts and adapts to an individual's physiological and cognitive state and reflects this via carefully designed avatars, environments, and interactions. I will discuss this concept in terms of three research directions: bioresponsive avatars, mediated communications and assistive tools. A preprint draft of this concept can be found on the following link: (`https://doi.org/10.48550/arXiv.2311.16610`)

3.13 Social VR for Social Skills Training

Sylvia Xueni Pan (Goldsmiths, University of London, GB; x.pan@gold.ac.uk)

Main reference Marco Gillies, Xueni Pan: "Virtual reality for social skills training", pp. 83–92, 2019.
URL http://dx.doi.org/10.1255/vrar2018.ch9

My research is about using VR to make our real life better. I am interested in creating VR applications with virtual humans that can be applied in different areas such as training, therapy, and education. For instance, in our early work in 2007 we used a friendly virtual character to help participants practise their social skills, so they can build more confidence for real life social interactions. More recently we developed a few scenarios in the area of health and healthcare related communication skills training, including understanding the psychological impact of domestic violence for social workers. Another important aspect of my work to use Social VR to help us understand real-world social interactions, which then informs the future design of social VR. For instance, we collaborate with neuroscientists to design and implement experimental studies which contributed towards understanding the brain mechanism behind autism.

3.14 Goal-adaptive Collaborative Spatial Experiences with GenAI

Payod Panda (Microsoft Research – Cambridge, GB; payod.panda@microsoft.com)

The collaborative lifecycle involves more than meetings. Collaboration occurs at several timescales – the "work planning" timescale (days to weeks), the micro timescale (or "in-the-moment" interactions – scale of seconds), and the macro timescale (e.g., at the scale of projects – months to years). Additionally, effective collaboration is effortful, but traditional collaboration systems offer little support for reducing this effort across the collaborative lifecycle. For example, meetings often do not list what the goals of the meeting are, nor what is expected from meeting attendees. HCI has largely addressed the micro timescale of interactions – what kinds of interactions should a system provide in order to support collaborative tasks? We need to shift from designing for "moments" to designing for "workflows", which should be driven by collaborative *goals*. How could we assist users to transition between activities within and across timescales in order to accomplish their short-, mid-, and long-term goals? I propose using Generative AI (GenAI) systems in order to adapt the meeting interface to the individual, team, and organizational goals, involving interactions like reconfiguring a collaborative space (3D virtual environment) and rearranging task elements in the space (distribution of task space).

3.15 What Can Social XR Do for Us that Traditional Communication Technology Cannot, and How Can We Know?

Alexander Raake (Audiovisual Technology Group, I3TC – TU Ilmenau, DE; alexander.raake@tu-ilmenau.de)

Our team conducts research on audiovisual technology, perception and experience. A specific focus lies on telepresence technology used for human-to-human communication. We address Augmented, Virtual or Mixed Reality (AR/VR/MR, eXtended reality, XR), as well as robotics. Here, we integrate the multimedia-driven, initial approaches of Quality of Experience (QoE) assessment with the experience evaluation methods evolving in the AR/VR/MR community over decades, such as presence, social presence and co-presence, plausibility, or cybersickness. Besides direct evaluation methods using questionnaires, we employ indirect methods such as behavior and conversation analysis, regarding verbal and non-verbal communication. Here, the impact of non-obvious technical properties are of interest, such as that of transmission delay. In this case, quality and (audiovisual) fidelity may appear very high, but the individual temporal realities may be out of sync. In previous research, we showed that attribution may then be to the other person(s), not the system, e.g., considering the (previously unknown) other as less extrovert or open, when delay was on the line. Besides the visual modality, our group is interested in impact of audio and hearing, as well as audiovisual integration for attention, cognition and communication. Beyond the Social XR experience-perspective, we are interested in the "resources" involved, in terms of sustainability: (1) The human mental and physical resources spent, for example measuring fatigue for MR-based telepresence versus meeting face-to-face, or the positive impact on wellbeing with mediated social presence. (2) The amount of energy and natural resources consumed along the end-to-end chain (e.g., by a given media system implementation versus another), or resources saved (e.g., meeting via videoconferencing or MR rather than travelling). For the seminar, I would like to jointly specify a common set of research methods and use cases to be considered, aiming to address key challenges brought to the seminar. Moreover, I am interested in collaboratively elaborating a selection of these challenges and possibly develop initial ideas on how to address them.

3.16 Instrumenting for Understanding Social XR Experiences

David A. Shamma (Toyota Research Institute – Los Altos, US; aymans@acm.org)

Research on Social XR has seen two fronts. One is the exciting, far vision of the future inspired by design fiction narratives and imagining technology beyond our capacity to build. The other is what we can make with today's impressive but limited technology to adapt to tablets, web browsers, and head-mounted displays. Between these two is a field ripe for research because one can measure, test, and evaluate how people behave, interact, and enrich their lives with XR technology. As we step forward, our research should address theory-informed social conditions in the real world and explore how these patterns manifest in XR environments. It is not enough to instrument the virtual and augmented worlds. We should alter what is physically possible into the impossible, as XR's great potential lies in creating non-realistic experiences. These unreal XR experiences have the ultimate potential to unlock stronger interactions and collaborations, and they should require exploration as technology takes each step forward.

3.17 Bringing Real People into XR

Aljosa Smolic (Lucerne University of Applied Sciences and Arts – Rotkreuz, CH; aljosa.smolic@hslu.ch)

Social XR inherently requires digital representations of humans. For the visual part this is some kind of 3D computer graphics model. In most XR applications today we find purely computer-generated models which may be referred to as avatars. As an alternative it is possible to reconstruct 3D models of from images and video by means of 3D computer vision. The result is often referred to as volumetric video/holograms (VV). While many aspects of VV technology from capture to display have reached a high level of maturity, still a lot of problems remain unresolved to make it widely acceptable for social XR and telepresence applications.

3.18 Requirements for Future Social XR Applications

Anthony Steed (Department of Computer Science, University College London, GB; A.Steed@ucl.ac.uk)

My research started out in the low-level engineering of collaborative virtual reality systems. In my talk I started by presenting some of our early work on social VR applications, and how, at the time, the main problems with the graphics and network engineering. I then presented some more recent work that focuses on identifying the key technical challenges for systems that support effective communication. For example the role of eye-tracking, body-tracking and latency.

Our previous work focused on technical demonstrations and lab-based experiments. Going forward, we are trying to do more longitudinal studies of social XR use to identify how users adapt after time. We are very interested in building social applications that support users with different literacies and competencies with VR.

Finally I talked about and demonstrated our Ubiq toolkit. Ubiq supports a variety of AR and XR devices. Client APIs, demonstrations and server code is completely open source so that anyone can set up secure, GDPR-compliant social systems. We talked about how we recently extended it to support instrumentation and scalability. We built a Virtual Dagstuhl social VR demo.

3.19 Adaptive Social XR

Kashyap Todi (Meta – Redmond, US; kashyap.todi@gmail.com)

As a research scientist at RL-R, I work at the intersection of Human–Computer Interaction (HCI) and Artificial Intelligence (AI) towards solving emergent XR interaction problems. My expertise and interests are mainly around applying computational methods to address core HCI problems systematically. I have been doing so in domains of generating user interfaces via models of interactions and adapting user interfaces to individual users and their context. I believe that contextually adaptive UIs and interactions will be critical for enabling highly performant and usable XR applications and experiences. This will require extensive research on key components including modeling users, environments, and interactions, developing AI and/or computational approaches for optimization and adaptation, collecting and formatting extensive training datasets, identifying highly reliable quantitative metrics and evaluation methods, and finally close alignment with end-user applications. While I've worked extensively on "Solo-XR" scenarios in my research so far, I believe this philosophy and approach will be crucial and beneficial for Social XR settings as well. As such, I encourage and urge everyone to consider: *what might adaptive social XR look and feel like in the future?*

3.20 Human-centric Factors in Immersive Communication

Irene Viola (Centrum Wiskunde en Informatica – Amsterdam, NL; irene.viola@cwi.nl)

My research relates to quality of experience in immersive multimedia systems. In particular, I am interested in understanding the user at the center of immersive systems: how do they behave, what they are interested in, how do they interact with each other and with the media objects. There are some core challenges in how we measure the user experience, whether it is qualitative or quantitative, explicit or implicit; how we can predict the reaction of the user in such experiences, whether it is the way they will move, what they will focus on, or whether they'll want to replicate the experience; and how we can use such measurements and predictions to optimize the system and make it user-centric.

3.21 Fostering Well-being, Communication & Empathy with XR

Nadine Wagener (University of Bremen, DE; nwagener@uni-bremen.de)

In my research I explore how to design technology fostering well-being, mental health, communication and empathy with XR. I explore two main aspects: 1. How can a XR system support a self-exploration approach of own emotions and our "inner worlds", e.g. by offering passive haptic feedback and prompts to facilitate self-awareness and self-reflection?, and 2. How can we share (and collaboratively explore) these "inner worlds" with a social ecosystem? As one example, users can invite friends or colleagues into VR spaces that they autonomously create to represent their emotions (e.g. in regard to a shared conflict), and can also collaboratively express their emotions through art in VR. This approach can provide the foundation for developing a shared language, mutual understanding, and effective conflict management. I am further interested in including physiological data to make inner states accessible to oneself and others, focusing on finding means to represent those data in a qualitative way, e.g. through mirroring stress with a VR thunderstorm. I look forward to discussing different modes of biosignal inclusion and ethics in regard to SocialXR.

3.22 AI-driven 3D Digital Humans in XR

Zerrin Yumak (Utrecht University, NL; z.yumak@uu.nl)

My research is about 3D digital human technologies in games and Social XR applications. In particular, I am focusing on AI-driven non-verbal behavior synthesis algorithms for facial expressions, gestures and gaze behavior using deep learning algorithms. The goal of the research is to automatically generate animations conditioned for instance on audio and text and to create convincing and believable animations. That is useful to support the costly game development pipelines as well as for interactive applications where motions need to be generated on-the-fly. My work is data-driven and Motion Capture and Virtual Reality Lab at Utrecht University becomes instrumental for collecting human movement data for my research. I presented an overview of our research work during the seminar including FaceXHubert and FaceDiffuser. Another aspect of my research is the perception of animations to better understand what aspects of these characters makes them accepted by users. I have also done research on socially interactive characters in particular on the topics of emotion and memory modeling and multi-party interaction. The connection between AI, XR and HCI is the core of my research which is also discussed in our IEEE VR MASSXR Workshop.

4 TUESDAY Working Groups

During the Tuesday session, we set the challenge of "Co-envisioning Social XR futures and how we may achieve them", breaking researchers into four working groups. Working groups used shared Miro boards to brainstorm key concerns and organize them into output summaries, shared below. When considering possible futures, participants were asked to identify key assumptions and values, as well as tools that will be needed, alongside their visions.

4.1 Group A: Synchronizing Asymmetric Individual & Shared User Perspectives in XR

Linda Hirsch (LMU Munich, DE)
Katherine Isbister (University of Santa Cruz, US)
Payod Panda (Microsoft Research, Cambridge, UK)
David Ayman Shamma (Centrum Wiskunde & Informatica , Amsterdam, NL)
Kashyap Todi (Meta, Redmond, WA, US)
Zerrin Yumak (Utrecht University, Utrecht, NL)

Envisioning future developments of social XR, we imagine that asymmetric social experiences will become ubiquitous, easy to use, and integrated into daily routines. This change toward blended realities introduces multiple challenges. One challenge questions the balance between personalization and individual interests versus shared understanding and social experiences.

The increase in personal devices and possibilities to customize your virtual environment to create one's own "reality" opposes the idea of social XR experiences. In comparison, social experiences require a shared understanding of shared activities, a common language, or consented social norms and practices. In the future, we expect a multiverse of XR realities, including "My reality", "your reality" and "our reality". Such a multiverse questions privacy settings and the degree to which realities blend. For example, would putting flowers on someone else's table in VR impact the physical private household, and to what extent? Similarly, we assume we will have multiple virtual proxies interacting with others in meetings or social events for us. Proxies should be distinguished between passive and active, with passive being virtually present without further interaction and active being interactive. Interacting with a proxy instead of a "real user" will challenge social norms and relationships, which will have to be observed in the future. Additionally, **being present at all times** through proxies is the next step to **being available at all times**. Current issues of overloading users and induced stress are already consequences, which will potentially worsen. Thus, an important next step is to balance users' attention and set their well-being into focus. We summarize inventions that will drive and challenge future social XR: AI & Proxies, Adaptive Interfaces, Multi-Device Constellations, Multi-Location Constellations, and Privacy.

4.2 Group B: What Kind of XR Future Do We Hope to Have? (Or Rather: "What Kind of Aspects Do We Foresee to be Relevant for an XR Future"?)

Eric J. Gonzalez (Google, Seattle, US)
Josh Greenberg (Alfred P. Sloan Foundation, New York, US)
Jie Li (EPAM, Hoofddorp, NL)
Alexander Raake (Audiovisual Technology Group, TU Ilmenau, DE)
Aljosa Smolic (Hochschule Luzern, Rotkreuz, CH)

We started with broad brainstorming, then clustered responses across a broad spectrum from technical (see Fig. 3, left) to overall goals and ethical considerations (middle) and super-powers (right). Applications of more general nature are on the bottom left, more to the right commercial applications. An aspect discussed for some longer time is that of feature-access control, based on desired privacy and properties of bi-lateral / multi-lateral relationships (e.g., share photorealistic info only with dedicated others). Can access to such info be controlled asymmetrically? Will the space be one world or a sort of multiverse where persons can be in instances of the same space?

Figure 3 Overview of results Group B, What kind of XR future do we hope to have? (Or rather: "What kind of aspects do we foresee to be relevant for an XR future"?).

4.3 Group C: Redefining Common Grounds in Social XR

Sun Joo (Grace) Ahn (University of Georgia, Athens, US)
M. Doga Dogan (MIT, Cambridge, US)
Jan Gugenheimer (TU Darmstadt, DE & Telecom Paris, FR)
Yun Suen Pai (The University of Auckland, NZ)
Sylvia Xueni Pan (Goldsmiths, University of London, GB)

We discussed how XR platforms today do not share a common ground; a "ground truth" for a society to function properly. Fragmentation of media platforms has led to a reduction and loss of common grounds. This creates difficulty in determining the agenda or importance of problems that must be solved (e.g., war? climate change? gender issues). When there is no common ground, who determines what the ground truth is? How do we determine what is normal and what is not? How do we prepare against misinformation/deepfakes? These discussion points are illustrated in Figure 4. We propose that social VR can be used to re-establish common ground via the following mechanisms:

- Allow companies to publish their core values. Users will select which social XR service they want to engage in.
- Use social XR and immersive experiences to highlight important problems and solutions (e.g., globalize local events)
- Social XR researchers will need to learn how to 'break' the experience first so that we can prepare safety/protective tools (we can figure out what can go wrong)
- Users' ability to determine the credibility of the information source will be important

Figure 4 Overview of results for Group C's discussion regarding the lost of common grounds in XR.

4.4 Group D: Development and Implementation of Social XR Systems

Kai Kunze (Keio University, Yokohama, JP)
Neil McDonnell (University of Glasgow, GB)
Anthony Steed (University College London, GB)
Irene Viola (Centrum Wiskunde en Informatica, Amsterdam, NL)
Nadine Wagener (Universität Bremen, DE)

The group explored various aspects of Social VR system development, raising concerns and considerations from different angles. We emphasized the challenges of creating user-friendly, data-protected, and persistent VR systems, while focusing on the balance between avatar personalization and privacy. Data control, privacy, and potential attention manipulation through eye tracking in VR scenes were discussed, highlighting the need for ethical considerations.

Opportunities and threats associated with making internal physiological data visible in XR were explored, considering individual willingness to share such personal information. The challenges and ethical concerns in XR, particularly personalized advertising and societal behavior impact, were highlighted. Topics included the fragmentation of reality in XR, the potential negative impact on communities, and the balance between individualism and commonality.

The discussion touched on the societal implications of allowing individuals to curate their XR environments. Concerns about data exploitation in XR, issues with GDPR enforcement, and the need for responsible data management were expressed. The difficulty of conducting experiments with social XR platforms due to data privacy issues and GDPR bureaucracy was discussed.

The discussion concluded with a focus on designing future XR spaces with consideration for minimizing exploitative practices, keeping data local, and ensuring accountability for companies involved in creating XR experiences. Overall, the discussion underscored the complex ethical and practical considerations in the development and implementation of Social VR systems.

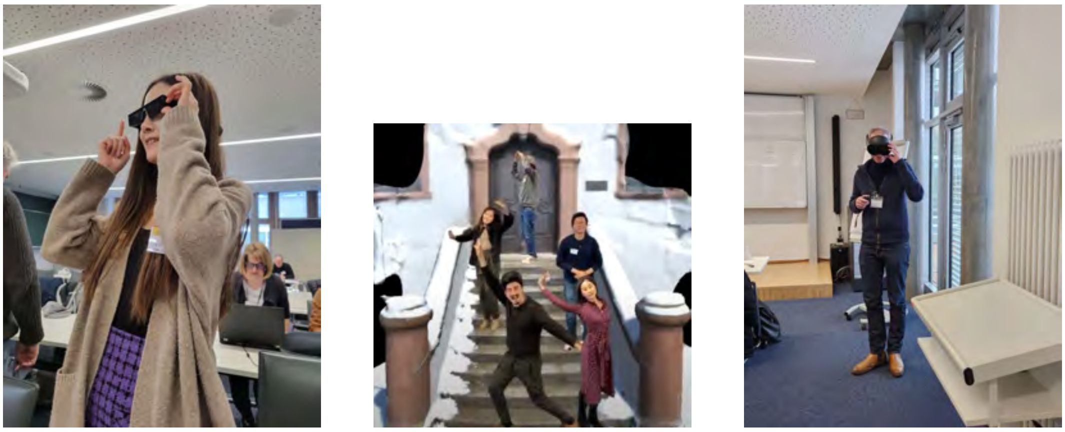


Figure 5 Participants try out different tools and hardware for social XR. **Left:** a participant tries out AR glasses. **Center:** a screenshot of a video that has compiled five separate Vologram captures of participants dancing, superimposed onto the Dagstuhl steps. **Right:** a participant trying out Ubiq in virtual reality.

Figure 6 Participants try out the Embodied Telepresent Connection (ETC) demo.

5 TUESDAY Demos

On Tuesday afternoon, participants were invited to share demonstrations of their research related to social XR. We tried out some of the latest social XR hardware (see Figure 5: left). While other participants presented several different software solutions including for supporting social connection (see Figure 6) and avatar expression, as well as tools for networking (see Figures 5: center and 2: right) and recording XR content (see Figures 5: right and 7):

- Alexandra Kitson – Embodied Telepresent Connection: Interactive installation to support connection and pseudo-touch `http://ispace.iat.sfu.ca/project/etc/`
- Alijosa Smolic – Volograms: record a video and turn it into an AR experience `https://www.volograms.com/`
- Anthony Steed – Ubiq: a free, open-source networking library for research, teaching and development `https://ubiq.online/`
- Zerrin Yumak – FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis using Self-Supervised Speech Representation Learning `https://github.com/galib360/FaceXHuBERT`

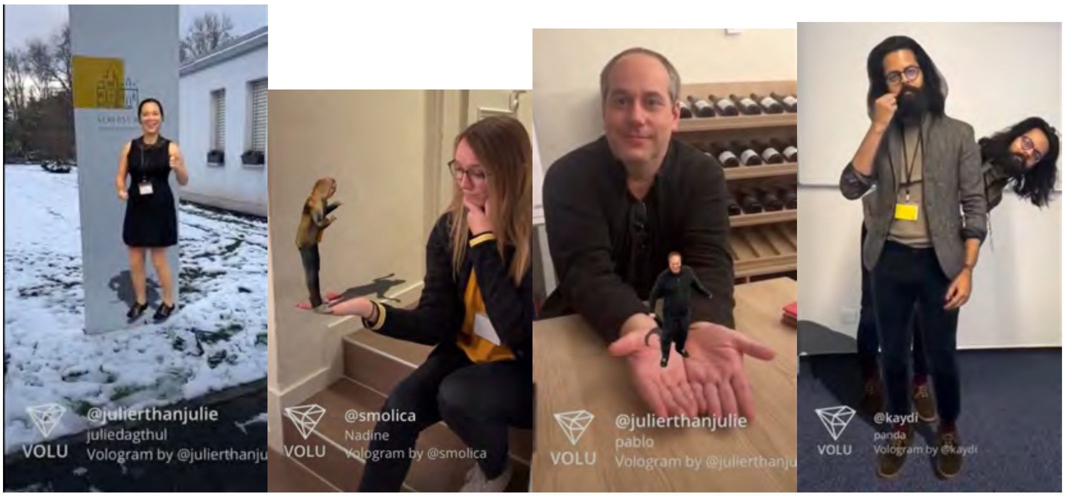


Figure 7 Participants try out the Vologram Capturing tool and experiment with creating AR volograms with their smart phones.

6 WEDNESDAY Working Groups

The Capturing and Modeling session explored the future of capturing and modeling people and places.

It began by discussing the history of place capture, from early attempts in the 1990s to more recent advances like Microsoft's Holoportation or Avatar Codec. The presentation then looked at the future of people capture, with a focus on high-quality streamable free-viewpoint video and the rise of virtual humans (or "vTubers").

The session then discussed some of the challenges and ethical implications of capturing and modeling people and places. It asks thought-provoking questions about what communication cues should be captured and shared, what elements of the environment should be shared, and how we can separate out communication cues from representation. The presentation also discussed the potential long-term social impact of this technology, and asks questions about how we prepare people for the additional cognition needed to deal with the virtual and AI worlds.

The conclusion was a hands-on task in which participants were asked to build and test many of the tools presented and reflect about their own experiences with capturing and sharing place and people. They were asked to consider how difficult it was to capture place and people with current tools, how effective the current captured content is for communication, and what would need to happen for their grandparents to capture people and place.

The session was meant to be a thought-provoking and engaging exploration of the future of capturing and modeling people and places. It raised important questions about the ethical implications of this technology and encourages participants to think critically about their own experiences with capturing and sharing place and people.

A good snip of the creations from this session is available on: `https://x.com/twi_mar/status/1730062159151800399?s=20`.

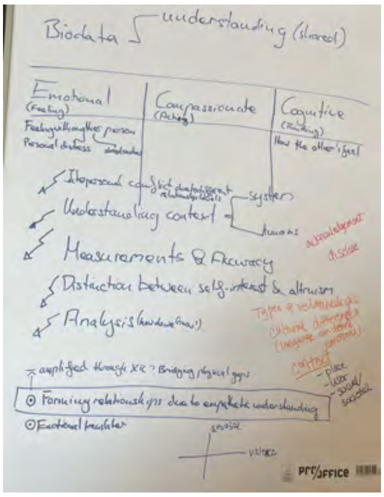


Figure 8 Notes on empathic computing from Thursday Working Group A taken in the News Room at Dagstuhl.

7 THURSDAY Working Groups

7.1 Group A: Empathic Computing

Sun Joo (Grace) Ahn (University of Georgia, Athens, US)
Mark Billinghurst (University of South Australia, Adelaide, AU)
Linda Hirsch (LMU München, DE)
Alexandra Kitson (Simon Fraser University, Surrey, CA)
Yun Suen Pai (The University of Auckland, Auckland, NZ)
Nadine Wagener (Universität Bremen, DE)

Our working group was tasked with answering the following three questions: What theories should be driving this research? How can we incorporate empathic computing into our communications? Long term impact of engaging with empathic experiences? Figure 8 documents some of our notes and discussion.

7.1.1 Relevant Theories and Conceptualization of Empathy

We discussed different models of empathy, recognizing that there is not a consensus among the scientific community around the precise definition of empathy and its constructs.

Goleman & Ekman [1]: Three dimensions of empathy (cognitive, emotional, compassion)

1. Cognitive refers to perspective taking (I understand you)
2. Emotional refers to feeling personal distress and sharing emotions (I feel you)
3. Compassion refers to actionable outcomes (I help you)

Davis [2]: Individual differences in empathy (perspective taking, fantasy, empathic concern, personal distress)

- Measured through Interpersonal Reactivity Index (IRI)

Batson et al. [3]: Altruism vs. self-interest

- What is the motivation of empathy?

Based on the above theories and our discussion, we formed a working definition of empathic computing: *Forming meaningful relationships in social XR – empathic computing may provide the foundations for meaningful social XR.*

7.1.2 Current Challenges of Empathic Computing

Second, we discussed the current challenges of incorporating empathic computing into our communications.

- Objective vs. Subjective assessment
- Defining and conceptualizing empathy
- Systems struggle with the multi-layered complexities of context (e.g., place, user differences, social relationships)
- Physiological Signals are also very context dependent
- Can AI "train" people to become more empathic?
- Accessibility and usability of wearables and sensors in empathic systems, including problems of scaling
- How do you express micro-cues to facilitate communication (e.g., turn-taking)

7.1.3 Empathic Computing in Social XR

Next, we narrowed in on two main challenges of incorporating empathic computing into our communications: context aware XR and representing emotions in XR:

Context-Aware XR: The necessary contexts for empathy include Place, User, Social, Past Experience, Relationships, and Systems that acknowledge its limitations and disclose learning process (transparency).

Representing emotions in XR: Some of the ways include Avatars, (objects in) the VE, and separate virtual entities.

7.1.4 Long-term Impact

Finally, we discussed some of the long-term impacts of empathic computing in social XR.

- Interpersonal conflict due to different relationship intimacy between users and empathic systems
- Privacy concerns: the dilemma between privacy vs. context-aware. For example, users want to understand others' emotions; don't want to reveal theirs
- Ethical concerns related to long-term relationships with empathic systems
- Concerns related to over-gamified systems (can you gamify relationships)
- Potential risk of hyper-empathy (caring too much)

7.1.5 Grand Challenges

Scientific Evaluation of Empathy. Measuring empathy and making it comparable is a continuous challenge because emotions are highly subjective. Research increasingly complements qualitative data with quantitative biodata measurements. For social XR, we need an understanding on an individual and collective level. Thus, we suggest a focus on mixed methods approaches for the scientific evaluation of empathy and see benefits in supporting AI models that take into account the user(s)'s user traits, background, etc.

Ethical Concerns. Virtual technologies allow customization to a great extent based on how users feel and their preferences. Empathic computing provides the data and technological setup to implement this for individual and social contexts. However, sharing biodata on this level raises privacy concerns, which can easily be taken advantage of. Similarly, empathic computing can support understanding others' emotions. However, this also means that we hand over our empathy to a system to tell us, in return, about how another person feels. This requires great trust in an empathic system, further requiring transparency and a certain level of user control.

Furthermore, empathic computing increases the lack of transparency about users' intentions when engaging in social interaction or showing empathy. Sometimes, this will be beneficial, such as when a doctor talks to patients about their diagnosis. However, in a more intimate relationship, fake or pretended empathy is not sustainable for a healthy relationship. It questions how we can disclose user intentions and raise system transparency for protecting misuse and contributing to "good" social relationships and connections.

Context Adaptation & Reaction. Emotions are context- and person-dependent. This requires empathic computing integrated into social XR to be context-aware and -sensitive when gathering, evaluating, processing, and displaying the data. Challenges arise on different levels. One relates to training models on being context-sensitive and having a comparable data set over multiple situations. Another derives from a user perspective where emotions are expressed very differently for different purposes depending on the cultural background. This also leads to the system's sensitivity regarding when to disclose a user's emotions to others and when not.

Evolving XR. One of the grand challenges is the anticipation of how social XR will evolve over generations of users and technological advancements. It might lead to a greater digital divide between users of social XR and non-users, followed by a diverging understanding of (social) reality. Similarly, AI systems and proxies will become regular members of our social system, communication, and collaboration. Yet, it is currently not predictable to what extent, in what ways, and how we, as human users, can keep control.

References

1 Daniel Goleman, *The brain and emotional intelligence: New insights*, More than sound Northampton, MA, 94, 47–48, 2011.
2 Mark H Davis, *A multidimensional approach to individual differences in empathy*, American Psychological Association Washington, DC, 1980.
3 C Daniel Batson; Janine L Dyck; Randall J Brandt; Judy G Batson; Anne L Powell; Rosalie M McMaster; Cari Griffitt, *Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis*, Journal of personality and social psychology, American Psychological Association, 55:1, 1–52, 1988.

7.2 Group B: Blended Reality

M. Doga Dogan (MIT, Cambridge, US)
Eric J. Gonzalez (Google, Seattle, US)
Katherine Isbister (University of Santa Cruz, US)
Kai Kunze (Keio University, Yokohama, JP)
Payod Panda (Microsoft Research, Cambridge, GB)
Sylvia Xueni Pan (Goldsmiths, University of London, GB)

7.2.1 Multi-sensory Social Experiences

Much advancement in the area of Virtual and Augmented reality has been mainly driven by our ability to push the boundary of computer graphics and audio display. Real-life experiences does not stop at what we see and what we hear. One of the challenges in the area of creating effective multi-sensory social experiences is the simulation and display of senses beyond visual and audio, such as social touch. There are few reasons behind this challenge:

- Those senses could plays a bit part in shared experience, for instance, the small (olfactory) of mulled wine and roast chestnuts could people of the Christmas Market experience. However, as they are normally experienced at a **subconscious level**, it is difficult to describe and come up with a set of rules to code for in a simulated environment.
- **Device challenge** – using haptics as an example, there is no general device that could address the richness of touch (e.g., tactile, weight, pressure). This is also a big challenge in making a business case for a particular type of haptics in a consumer device.
- **Individual differences** in our perceptual threshold: our sensitivity in distinguishing different weights, temperature could be quite different; making it very difficult to control the experience with pre-defined code. There is also the challenge of culture differences – if someone from Japan were to greet someone from France, should the Japanese bow be translated into a hug and two kisses on the cheek? Shall we introduce asymmetric social interaction to calibrate the social experience, or should we maintain the authenticity at the cost of creating a potentially very awkward social interaction?

7.2.2 Semantic Understanding of Physical and Social Context

To effectively blend real and virtual environments, objects, and experiences, XR systems must have a rich understanding of the physical and virtual worlds. This includes having a sense of which objects in the user can touch or interact with (e.g., a desktop surface) and which should be avoided (e.g., a glass of water on the desk). Not only does this make the system "smarter" and allow it to provide information that is more relevant to the user's surroundings, it can also improve interaction affordances by opportunistically aligning real and virtual content for improved haptic experiences (e.g., displaying touch UI on tabletop). The same concepts extend to social contexts, where it may be beneficial to alter how certain XR information is displayed depending on, for example, whether the user is in private, having a conversation, driving, or in a public setting. Enabling this richer understanding by leveraging AI and computer vision tools will be essential for this task.

7.3 Group C: Assets and Datasets

Jan Gugenheimer (TU Darmstadt, DE & Telecom Paris, FR)
David Ayman Shamma (Centrum Wiskunde & Informatica, Amsterdam, NL)
Aljosa Smolic (Hochschule Luzern, Rotkreuz, CH)
Kashyap Todi (Meta, Redmond, US)
Zerrin Yumak (Utrecht University, NL)

7.3.1 How Can We Recreate "Social XR @ Dagstuhl 2023"?

During the week, we have collected and/or created various bits and bytes of data that capture varying aspects of this seminar. This includes:
- 3D Objects & Scenes
- Volumetric Video
- Images
- Audio
- Video

There are some immediate technical challenges that we can observe around data collection and use:
- Formats: interoperability, open vs. closed, etc.
- Scalability: quality, quantity, compression, etc.
- Metadata: additional info, synchronization, ...
- License (cc by-nc-nd?)

But the looming big question: how do we really capture and recreate the event experience? How does someone reliving this Dagstuhl understand and grasp "What happens at 20:00?" (spoiler: cheese plates are served in the wine cellar).

7.3.2 Beyond Space for Social XR Capture

We are missing data and information that might be essential for adequately capturing the event:
- Time (program/schedule/events), people, speech, emotional states, reactions, routines, experience.
- What is needed to recreate complex social organization? (Two teams occupying same space)

Coming out of these discussions, we have identified three grand challenges around assets and datasets for social XR, summarized in subsection 7.3

7.3.3 Grand Challenges

Social Stitching. Can a sparse set of assets be stitched together with AI to fabricate missing parts? What kind of quantitative and qualitative evaluation would be required? For this challenge, a sparse collection of a scanned space is distributed. The goal is to make a cohesive scene/world. For the Dagstuhl assets, one would need to include a program (meals, breaks, seminar schedule, etc.) and floor prints (as seen around the building). A scan of a few guest rooms would also be good to add. Challengers would stitch the assets together to make Dagstuhl. Tools like GenAI could be used to hallucinate or speculate what's in each room, who might be occupying the room, and what people are doing in the rooms. This would have two evaluations: quantitative and qualitative. Both would be evaluated given a program and personnel load. Evaluators should be able to drop into (or jump to) any time and feel the scene.

- For the quantitative, one would need a high res scan of a few rooms and we could measure accuracy.
- For qualitative, it would have to feel like Dagstuhl.

An AR variant of the stitching challenge could be an additional evaluation metric.

Authoring and Sense-making. What approaches can enable sense-making and authoring to reconstruct the spirit of the event? How can pieces in the form of different assets (3D, audio, video, text, etc.) be put together (authoring) to create a whole "event"? For that, what is a "model" of an event (space, time, humans, objects, audio, temperature, interrelations, etc.), what belongs to it in terms of abstract types? Given a (sparse) set of assets, related to an "event model", sampled from an event

- how can we interpolate and complete it, i.e. "make sense"?
- How can this go beyond immediate asset types that can be captured (e.g. video, 3D), but have to be inferred/derived (e.g. mental state of a person, interrelations).
- Can those be represented by some kind of network of nodes?

Preserving Privacy. How can privacy be preserved given the increasing fidelity of capture devices? What is privacy? What needs to be "preserved"? Risks and opportunities when embedding/representing data in latent spaces? What hardware alternative to RGB cameras could be used to capture environments that might be inherently privacy preserving (e.g., Lidar ?) Our understanding of "private" data might need to be extended beyond the captured data to also incorporate inferenced data from large reconstructions (e.g., can I use a full recording of the Dagstuhl event to figure out what Kash's favorite food is?)

7.4 Group D: Survey Instruments

Jie Li (EPAM – Hoofddorp, NL)
Sylvia Xueni Pan (Goldsmiths, University of London, GB)
Alexander Raake (Audiovisual Technology Group, TU Ilmenau, DE)
Anthony Steed (University College London, GB)
Irene Viola (Centrum Wiskunde en Informatica, Amsterdam, NL)

7.4.1 Tension between Transparency and Superpowers

The tension between transparency and the desire for social superpowers is a central challenge in the context of social XR. On one hand, the pursuit of social transparency can be a way to effectively create a medium in which people interact and communicate in a way that is indistinguishable from real life. This involves supporting situations that are already socially intricate, such as brainstorming sessions, and ensuring support for accessibility, with a focus on the clarity of identity. The acknowledgment of existing biases, such as the proteus effect, underscores the need for careful consideration – would we need to replicate such effects in a transparent social VR medium? Is it desirable? Usability is emphasized as a crucial factor, aiming to prevent any loss of information. Transparency is certainly important for very broad accessibility, such as bringing in family, professional or support groups with social dynamics that are important to preserve. For example, assisting someone undertake a task in a virtual world that is an extension of a situation in the real world (e.g. some forms of training or support).

On the other hand, the collective interest in social "superpowers" is evident, as individuals seek to don "mask" personalities in novel situations. However, the attractiveness of these situations may vary depending on individual personalities. The intention is to replace or enrich experiences to re-empower diminished cues, recognizing the potential risk of "after-effects." A key point is that individual superpowers might interfere with others, or create an equivalent of filter bubbles.

Thus when taking a step back, there might be a notion of a collective superpower that is itself well understood by participants. For example, empathy-enhancing powers might be very desirable if everyone agrees that this is acceptable, and no-one person has special insight. A very specific type of collective superpower might be to create "equity" (for example, the Altspace eye-line normalization?). That is, the idea that any biases that one might be inherent in the real world (e.g. simply height) are reduced or removed, to create a fairer interaction space. While this could be interpreted negatively at a individual level, the idea would be that collectively this would be advantageous. This then leads to interesting questions of the role of anonymity or social conventions in these situations (c.f. discussions held in a "Chatham House" style, or simply the purpose of masking in programs such as "The Masked Singer").

This intricate balance between transparency and superpowers thus poses significant conceptual and practical challenges in the realm of social XR.

7.4.2 How to Achieve a Shared Platform?

The second grand challenge lies in devising a shared platform that facilitates collaborative recording, replaying, and immersive experiences. Instruments must be crafted to analyze recorded formats, ensuring both reusability and auditability. The platform must seamlessly support motion and behavior capture, alongside accommodating video for non-recordable volumetric scenarios. Drawing inspiration from tools utilized in Computer-Supported Co-operative Work (CSCW), challenges include incorporating features like video annotation and time-series analysis. Overcoming obstacles, such as navigating funding schemes and addressing the apparent need for individualized recognition, presents a distinctive challenge within the collaborative landscape of Social XR.

7.4.3 Is There a Missing Instrument Such As a Questionnaire?

The absence of an optimal questionnaire poses a significant challenge. For example, the sensitivity of social presence to prior familiarity underscores the need for tailored assessment tools. For unfamiliar participants, the focus may shift towards gauging how much they learn about others, while for familiar participants, understanding mutual comprehension becomes a key interest. Moreover, certain concepts that were predominant A potential avenue for questionnaire development involves reverting to the "usability" concept in social systems, building upon established frameworks like the SUS questionnaire. The discourse extends to the validity and relevance of "presence" in mixed and augmented reality, questioning the balance between usability and utility in virtual reality applications and considering the likelihood of repeated use for specific purposes.

Conclusions drawn from these considerations emphasize the identification of a set of constructs deemed crucial for Social XR. Addressing the topic of scenarios, a proposed strategy involves breaking down activities into modular components across space and time, creating encounters with specific affordances. This approach aims to enhance comparability between systems and contribute to a more nuanced understanding of Social XR dynamics.

Participants

- Sun Joo (Grace) Ahn
University of Georgia – Athens, US
- Mark Billinghurst
University of South Australia – Adelaide, AU
- Pablo Cesar
CWI – Amsterdam, NL
- Mustafa Doga Dogan
MIT – Cambridge, US
- Eric J Gonzalez
Google – Seattle, US
- Mar Gonzalez-Franco
Google – Seattle, US
- Josh Greenberg
Alfred P. Sloan Foundation – New York, US
- Jan Gugenheimer
TU Darmstadt, DE
- Linda Hirsch
LMU München, DE
- Katherine Isbister
University of California at Santa Cruz, US
- Alexandra Kitson
Simon Fraser University – Surrey, CA
- Kai Kunze
Keio University – Yokohama, JP
- Jie Li
EPAM – Hoofddorp, NL
- Neil McDonnell
University of Glasgow, GB
- Yun Suen Pai
Keio University – Yokohama, JP
- Sylvia Xueni Pan
University of London, GB
- Payod Panda
Microsoft Research – Cambridge, GB
- Alexander Raake
TU Ilmenau, DE
- David Ayman Shamma
Toyota Research Institute – Los Altos, US
- Aljosa Smolic
Hochschule Luzern – Rotkreuz, CH
- Anthony Steed
University College London, GB
- Kashyap Todi
Meta Reality Labs – Redmond, US
- Irene Viola
CWI – Amsterdam, NL
- Nadine Wagener
Universität Bremen, DE
- Julie Williamson
University of Glasgow, GB
- Zerrin Yumak
Utrecht University, NL