# DAGSTUHL REPORTS

**Volume 12, Issue 3, March 2022**

Tensor Computations: Applications and Optimization (Dagstuhl Seminar 22101)
*Paolo Bientinesi, David Ham, Furong Huang, Paul H. J. Kelly, P. (Saday) Sadayappan, and Edward Stow* ........ 1

Computational Models of Human-Automated Vehicle Interaction (Dagstuhl Seminar 22102)
*Christian P. Janssen, Martin Baumann, Antti Oulasvirta, Shamsi Tamara Iqbal, and Luisa Heinrich* ........ 15

Database Indexing and Query Processing (Dagstuhl Seminar 22111)
*Renata Borovica-Gajic, Goetz Graefe, Allison Lee, Caetano Sauer, and Pinar Tözün* ........ 82

3D Morphable Models and Beyond (Dagstuhl Seminar 22121)
*James Gardner, Bernhard Egger, William Smith, Christian Theobalt, and Stefanie Wuhrer* ........ 97

Framing in Communication: From Theories to Computation (Dagstuhl Seminar 22131)
*Katarzyna Budzynska, Chris Reed, Manfred Stede, and Benno Stein* ........ 117

Graph Embeddings: Theory meets Practice (Dagstuhl Seminar 22132)
*Martin Grohe, Stephan Günnemann, Stefanie Jegelka, and Christopher Morris* ... 141

**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*
November, 2022

*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.12.3.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 22101

# Tensor Computations: Applications and Optimization

**Paolo Bientinesi*[1], David Ham*[2], Furong Huang*[3], Paul H. J. Kelly*[4], P. (Saday) Sadayappan*[5], and Edward Stow†[6]**

1 **University of Umeå, SE.** pauldj@cs.umu.se
2 **Imperial College London, GB.** david.ham@imperial.ac.uk
3 **University of Maryland – College Park, US.** furongh@cs.umd.edu
4 **Imperial College London, GB.** p.kelly@imperial.ac.uk
5 **University of Utah – Salt Lake City, US.** psaday@gmail.com
6 **Imperial College London, GB.** edward.stow16@imperial.ac.uk

**Abstract**

This report documents the program and the outcomes of Dagstuhl Seminar 22101 "Tensor Computations: Applications and Optimization". Tensors are higher dimensional analogs of matrices, and represent a key data abstraction for many applications in computational science and data science. Widely used shared infrastructure exists for linear algebra, while, in contrast, for tensor computations, there is no consensus on standard building blocks. This Dagstuhl Seminar aimed to bring together users, and performance optimization specialists, to build such foundations.

We present the abstracts of the 5 tutorials and 14 talks given. The working groups and their outcomes so far are then presented.

**Seminar** 06.–11. March, 2022 – http://www.dagstuhl.de/22101

**2012 ACM Subject Classification** General and reference → General literature; Computing methodologies → Linear algebra algorithms; Applied computing → Physical sciences and engineering; Mathematics of computing → Mathematical software performance; Mathematics of computing → Computations on matrices

**Keywords and phrases** Tensor, Optimisation, Linear Algebra, Compilers, Benchmark, Domain Specific Language

**Digital Object Identifier** 10.4230/DagRep.12.3.1

## 1 Executive Summary

*Paolo Bientinesi (University of Umeå, SE, pauldj@cs.umu.se)*
*David Ham (Imperial College London, GB, david.ham@imperial.ac.uk)*
*Furong Huang (University of Maryland, College Park, US, furongh@umd.edu)*
*Paul H. J. Kelly (Imperial College London, GB, p.kelly@imperial.ac.uk)*
*P. Sadayappan (University of Utah, Salt Lake City, US, saday@cs.utah.edu)*



Linear relationships between quantities are one of the most fundamental and pervasive phenomena in mathematics, science and computing. While matrices encode linear relationships between exactly two quantities, tensors are an abstraction representing linear relationships between multiple variables. Tensor computations therefore provide an abstract language for computations that span an enormous range of application domains, including machine learning, quantum information systems, simulations based on solving partial differential equations,

* Editor / Organizer
† Editorial Assistant / Collector

computational chemistry and beyond. The tensor abstraction enriches our understanding of the structure of computations, and exposes common challenges and solutions that cut across different research communities.

While the mathematics of tensors is well-developed and extensively applied across all of these applications and beyond, there is far less commonality in the software abstractions and tools deployed to execute tensor computations. This is in stark contrast to matrix computations, where common abstractions and stable interfaces have led to widely used tools that bring high performance to across diverse application domains.

This Seminar explored this challenge, and made significant progress towards establishing foundations for common implementations – embodying the substantial body of knowledge on high-performance tensor computation strategies in common software libraries and domain-specific program generation tools.

The Seminar began with five tutorial lectures, offered by the organisers in partnership with selected leading figures in some of the relevant communities. We began by mapping some of the diverse terminology. We then provided tutorials exposing the quantitative and qualitative diversity in how different communities use tensor computations – aiming to build a common understanding of key concepts, notations, and building blocks. We focused on the following application areas:

- Quantum physics and chemistry
- Mesh-based discretisations for solution of partial differential equations
- Machine learning.

The final tutorial reviewed the challenge of establishing unifying software tools, highlighting the enormous body of work that has been done within application areas.

The second phase of the Seminar consisted of more detailed presentations from the participants. These included motivating applications, but focusing on the fundamental computational workloads, methods, and performance challenges. Building on this, we also had contributions focused on implementation – low-level performance considerations, algorithmic proposals, compiler algorithms and compiler infrastructure.

In the third phase of the Seminar, we separated into three teams. One explored benchmarking and datasets, another made substantial progress on proof-of-concept implementation work to connecting the high-level Tensorly library for tensor decompositions in machine learning to a lower-level tensor-vector products – achieving considerable performance advantage. Finally there was also a major and continuing effort to define a common domain-specific language and compiler representation for tensor contractions that supports both high-level optimisations and the use of high-performance low-level libraries.

This 2021 seminar built on progress made at an earlier seminar with the same title, in March 2020 – which was very heavily impacted by the coronavirus pandemic. This seminar was also affected, to a lesser extent – with a reduced number of on-site participants, partly compensated by very useful engagement with researchers joining online, albeit from distant timezones.

This seminar benefited from broader engagement with application domains – partly as a result of the work that was done on tutorials – which we hope to publish in due course. It also benefited from deeper engagement with developers of high-performance building blocks. Finally, we initiated a new and continuing effort to define a common language and a common intermediate language for code generation tools.

## 2 Table of Contents

**Executive Summary**
*Paolo Bientinesi, David Ham, Furong Huang, Paul H. J. Kelly, and P. Sadayappan* 1

**Overview of Tutorials**

What is a tensor? What might a tensor abstraction look like?
*David Ham* . . . . . 5

Computing with tensors in mesh-based PDE discretisations
*Lawrence Mitchell* . . . . . 5

Tensors in Machine Learning
*Jeremy Cohen* . . . . . 5

Software for tensor computations: What is happening?
*Paolo Bientinesi* . . . . . 6

**Overview of Talks**

Infrastructure for Tensor Compilers: Lessons and Ongoing Developments from the MLIR Project
*Albert Cohen* . . . . . 6

Tensorly toolbox, and the tensoptly project
*Jeremy Cohen* . . . . . 6

Overview of the Boost.uBlas Tensor Extension
*Cem Bassoy* . . . . . 7

Sequences of tensor contractions: A design space exploration
*Carsten Uphoff* . . . . . 7

Hashing for the zeros/nonzeros of a sparse tensor
*Bora Ucar* . . . . . 7

Portable and efficient array redistribution
*Norman A. Rink* . . . . . 8

How will End-of-Moore impact high-performance tensor-centric applications?
*P. Sadayappan* . . . . . 8

Successes and Challenges for continuous matrix product states
*Jutho Haegeman* . . . . . 8

Simulation of the Sycamore quantum circuits with tensor networks
*Pan Zhang* . . . . . 9

Implicit Regularization in Deep Learning: Lessons Learned from Tensor Factorizations
*Nadav Cohen* . . . . . 9

The Tensor Brain: Semantic Decoding for Perception and Memory
*Volker Tresp* . . . . . 10

Matricized Tensor Times Khatri-Rao Product (MTTKRP)
*Christos Psarras* . . . . . 10

Domain-Extensible Compilers and Controllable Automation of Optimizations
*Thomas Koehler* . . . . . 10

**Working Groups**

TTV in Tensorly
*Jeremy Cohen, Cem Bassoy* . . . . . 11

A unified domain specific language for tensor contractions: Dagstuhl Tensor Language
*Teodoro F. Collin, Albert Cohen, David Ham, Paul H. J. Kelly, Thomas Koehler, Norman A. Rink, Edward Stow, Carsten Uphof, Sophia Vorderwuelbecke, and Connor Ward* . . . . . 12

**Participants** . . . . . 14

**Remote Participants** . . . . . 14

# 3 Overview of Tutorials

## 3.1 What is a tensor? What might a tensor abstraction look like?

*David Ham (Imperial College London, GB, david.ham@imperial.ac.uk)*



The 2020 Dagstuhl Seminar[1] spent a long time developing a common understanding of what tensors are. It also resulted in some ideas about multiple levels of abstraction for tensor computations. This presentation introduces these ideas and considers how different communities' expectations about tensors map onto them.

### References

1 Paolo Bientinesi, David Ham, Furong Huang, Paul H. J. Kelly, Christian Lengauer, and Saday Sadayappan. Tensor computations: Applications and optimization (dagstuhl seminar 20111), 2020.

## 3.2 Computing with tensors in mesh-based PDE discretisations

*Lawrence Mitchell (NVIDIA Corporation, Santa Clara, US, lmitchell@nvidia.com)*



At the core of residual assembly in finite element computations is a combination of tensor contractions and 'pointwise' non-linear operations. For efficient implementation, one often wishes to exploit structure in the tensors. I showed some structure that we use in the TSFC compiler [1, 2] to do this, and discussed some places where we might imagine using more generic technology.

### References

1 Miklós Homolya, Robert C. Kirby, and David A. Ham. Exposing and exploiting structure: optimal code generation for high-order finite element methods, 2017.
2 Miklós Homolya, Lawrence Mitchell, Fabio Luporini, and David A. Ham. TSFC: a structure-preserving form compiler. *SIAM Journal on Scientific Computing*, 40(3):C401–C428, 2018.

## 3.3 Tensors in Machine Learning

*Jeremy Cohen (CREATIS, CNRS, Villeurbanne, FR, jeremy.cohen@cnrs.fr)*
*Furong Huang (University of Maryland, College Park, US, furongh@umd.edu)*



Machine Learning is relying more and more on tensor computations. Tensors may represent extremely diverse data stemming from fluorescence spectroscopy, remote sensing, music information retrieval, image and video processing, but also parameters in high order statistics and deep learning. In this tutorial, after introducing several widely used tensor notations and low-rank approximation models, we detail how tensors and tensor models are used in several of these applications. In particular, we review the use of tensors in recommendation systems, blind source separation, dictionary learning, compression of neural networks (notably transformers) and classification. We finish this tutorial by pointing out a few important required tensor computations building blocks in the machine learning community.

### 3.4 Software for tensor computations: What is happening?

*Paolo Bientinesi (University of Umeå, SE, pauldj@cs.umu.se)*


**Main reference** Christos Psarras, Lars Karlsson, Paolo Bientinesi: "The landscape of software for tensor computations", CoRR, Vol. abs/2103.13756, 2021.
**URL** https://arxiv.org/abs/2103.13756

In stark contrast to the world of matrix computations, that of tensor computations still lacks a well defined set of building blocks. From an ongoing survey of the existing software for tensor computations, it emerged that many similar libraries are developed independently in different application domains. This inevitably leads to redundant effort and sub-optimal results (in terms of efficiency). In this talk we present and discuss possible building blocks to support high-performance tensor operations across different application domains.

## 4 Overview of Talks

### 4.1 Infrastructure for Tensor Compilers: Lessons and Ongoing Developments from the MLIR Project

*Albert Cohen (Google – Paris, FR, albertcohen@google.com)*



Peeking into MLIR for code generation and domain-specific compilation, with a focus on tensor algebra. Covering both graph-level and loop/vector-level optimization.

### 4.2 Tensorly toolbox, and the tensoptly project

*Jeremy Cohen (CREATIS, CNRS, Villeurbanne, FR, jeremy.cohen@cnrs.fr)*



The Tensorly toolbox is a high level library that provides tensor manipulations and decompositions in python[1]. After presenting the high level API and some recent additions from the Tensoptly project, we explore how bottleneck operations such as MTTKRP are implemented.

**References**

1 Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. Tensorly: Tensor learning in python. *Journal of Machine Learning Research (JMLR)*, 20(26), 2019.

## 4.3 Overview of the Boost.uBlas Tensor Extension

*Cem Bassoy (Fraunhofer IOSB, DE, cem.bassoy@iosb.fraunhofer.de)*



Boost.uBLAS has been recently extended with dense tensor types and basic tensor operations. In this talk interfaces and implementations of the extension are presented. Runtime results of tensor-time-vector and tensor-times-matrix operations are presented and discussed as well.

## 4.4 Sequences of tensor contractions: A design space exploration

*Carsten Uphoff (Intel Deutschland GmbH, Feldkirchen, DE, carsten.uphoff@intel.com)*



Kernels in finite element methods can be abstracted as sequences of tensor contractions. The number of implementation variants typically grows exponentially with the number of tensors. Sources of exponential growth are in particular the order of operations and the data structures of intermediate tensors. I discuss the design space of sequences of tensor contractions and algorithms to automatically select a fast implementation variant.

## 4.5 Hashing for the zeros/nonzeros of a sparse tensor

*Bora Ucar (CNRS and ENS Lyon, FR, bora.ucar@ens-lyon.fr)*


**Main reference** Jules Bertrand, Fanny Dufossé, Somesh Singh, Bora Uçar: "Algorithms and data structures for hyperedge queries", p. 28, 2022.
**URL** https://hal.inria.fr/hal-03127673

We consider the problem of querying the existence of hyperedges in hypergraphs. More formally, we are given a hypergraph, and we need to answer queries of the form "does the following set of vertices form a hyperedge in the given hypergraph?". Our aim is to set up data structures based on hashing to answer these queries as fast as possible. We discuss an adaptation of a well-known perfect hashing approach for the problem at hand. This is joint work with Jules Bertrand, Fanny Dufossé, and Somesh Singh and available as a technical report. There is also an efficient shared-memory parallel implementation [1].

### References

1 Somesh Singh and Bora Uçar. An Efficient Parallel Implementation of a Perfect Hashing Method for Hypergraphs. In *GrAPL 2022 – Workshop on Graphs, Architectures, Programming, and Learning*, pages 1–10, Lyon, France, May 2022. IEEE. to appear.

## 4.6 Portable and efficient array redistribution

*Norman A. Rink (DeepMind, London, GB, nrink@deepmind.com)*


**Main reference** Norman A. Rink, Adam Paszke, Dimitrios Vytiniotis, Georg Stefan Schmid: "Memory-efficient array redistribution through portable collective communication", CoRR, Vol. abs/2112.01075, 2021.
**URL** https://arxiv.org/abs/2112.01075

Computing on partitioned arrays in a distributed fashion has become commonplace, especially in the context of large-scale machine learning, where the size of large models necessitates working on partitioned arrays in order to fit into device memory. Computing on partitioned arrays typically requires communication in the form of redistributing chunks of arrays, which can easily become a performance bottleneck. I present a type-directed approach that synthesizes efficient communication sequences for array redistribution.

## 4.7 How will End-of-Moore impact high-performance tensor-centric applications?

*P. Sadayappan (University of Utah, Salt Lake City, US, saday@cs.utah.edu)*



The high-performance computing world has seen two "sea changes" so far. The first was the "attack of the killer micros" – change from vector supercomputers to highly parallel clusters of commodity processors. The next disruptive change was the emergence of GPUs with roughly an order-of-magnitude performance edge over CPUs for tensor computations. While tensor-centric applications in ML have largely adapted to this change, other domains are yet to benefit from the power of GPUs. It appears that the next disruptive change is looming with the end of Moore's Law scaling of VLSI. The latest breed of accelerators for ML all appear to be fully distributed-memory systems with thousands of processors on a chip without any shared-memory. The challenges in developing efficient tensor applications for these systems is even more daunting than GPUs. Compilers will need to play a significant role moving forward.

## 4.8 Successes and Challenges for continuous matrix product states

*Jutho Haegerman (University of Ghent, BE, Jutho.Haegeman@UGent.be)*


**Main reference** Benoît Tuybens, Jacopo De Nardis, Jutho Haegeman, Frank Verstraete: "Variational Optimization of Continuous Matrix Product States", Phys. Rev. Lett., Vol. 128, p. 020501, American Physical Society, 2022.
**URL** https://doi.org/10.1103/PhysRevLett.128.020501

A particular limit of the tensor train / matrix product state construction gives rise to a class of quantum states known as continuous matrix product states. The energy minimisation or dynamical evolution thereof gives rise to coupled set of non-linear matrix-valued partial differential equations. I discuss our successes in dealing with them so far, and the remaining challenges that we face for further applications.

## 4.9 Simulation of the Sycamore quantum circuits with tensor networks

*Pan Zhang (Chinese Academy of Science, Beijing, CN, panzhang@itp.ac.cn)*



The sampling problem of the Sycamore circuits has been used by Google to demonstrate the quantum advantage. I will introduce tensor network methods based on the sparse-state representation to generate one million uncorrelated samples from the final state of the Sycamore circuits with fidelity greater than Google's experiments.

## 4.10 Implicit Regularization in Deep Learning: Lessons Learned from Tensor Factorizations

*Nadav Cohen (Tel Aviv University, IL, cohennadav@cs.tau.ac.il)*


**Main reference** Noam Razin, Asaf Maman, Nadav Cohen: "Implicit Regularization in Tensor Factorization", in Proc. of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, Vol. 139, pp. 8913–8924, PMLR, 2021.
**URL** https://proceedings.mlr.press/v139/razin21a.html

The mysterious ability of deep neural networks to generalize is believed to stem from an implicit regularization, a tendency of gradient-based optimization to fit training data with predictors of low "complexity" [1, 2]. A major challenge in formalizing this intuition is that we lack measures of complexity that are both quantitative and capture the essence of data which admits generalization (images, audio, text, etc.). With an eye towards this challenge, I will present recent analyses of implicit regularization in tensor factorizations, equivalent to certain non-linear neural networks. Through dynamical characterizations, I will establish implicit regularization towards low rank, different from any type of norm minimization, in contrast to prior beliefs. Then, motivated by tensor rank capturing implicit regularization of non-linear neural networks, I will suggest it as a measure of complexity, and show that it stays extremely low when fitting standard datasets. This gives rise to the possibility of tensor rank explaining both implicit regularization of neural networks, and the properties of real-world data translating it to generalization.

### References

1 Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
2 Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

## 4.11 The Tensor Brain: Semantic Decoding for Perception and Memory

*Volker Tresp (Ludwig-Maximilians-Universität München & Siemens, München, DE, volker.tresp@siemens.com)*



I start with an introduction to PyKEEN [1], the main talk is on the tensor brain. It is a unified computational theory of an agent's perception and memory. In our model [2], perception, episodic and semantic memory are refined by different functional and operational modes of the oscillating interaction between index layer and a representation layer (global workspace) in a bi-layer tensor network (BTN).

**References**

1 Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. Pykeen 1.0: a python library for training and evaluating knowledge graph embeddings. *Journal of Machine Learning Research*, 22(82):1–6, 2021.
2 Volker Tresp, Sahand Sharifzadeh, and Dario Konopatzki. A model for perception and memory. In *Conference on Cognitive Computational Neuroscience*, 2019.

## 4.12 Matricized Tensor Times Khatri-Rao Product (MTTKRP)

*Christos Psarras (RWTH Aachen, DE, psarras@aices.rwth-aachen.de)*



Given the apparent interest at the start of the week on the MTTKRP operation, especially among the machine learning audiences, I decided to present the challengers and my insights on providing a generic black-box implementation of MTTKRP for 3D dense tensors and beyond. I presented a graph which listed the different ways of computing MTTKRP for 3D tensors, without explicitly transposing the underlying tensor, and emphasized on their (often significant) differences in performance. My findings pointed to the difficulty of creating a mechanism that can accurately predict which method for computing MTTKRP is most efficient depending on the target platform (CPU or GPU), the sizes of the dimensions of the underlying tensor, and the number of components (columns) of the factor matrices.

## 4.13 Domain-Extensible Compilers and Controllable Automation of Optimizations

*Thomas Koehler (University of Glasgow, GB, t.koehler.1@research.gla.ac.uk)*



I presented my work on a rewrite-based domain-extensible compiler with controllable automation of optimisations [1, 2, 3]. To encourage discussion, I presented two opinions on how future tensor compilers should be designed. “Will you agree, or have a different opinion?”

**References**

1 Bastian Hagedorn, Johannes Lenfers, Thomas Koehler, Xueying Qin, Sergei Gorlatch, and Michel Steuwer. Achieving high-performance the functional way: A functional pearl on expressing high-performance optimizations as rewrite strategies. *Proc. ACM Program. Lang.*, 4(ICFP), aug 2020.

2 Thomas Koehler and Michel Steuwer. Towards a domain-extensible compiler: Optimizing an image processing pipeline on mobile cpus. In *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 27–38, 2021.

3 Thomas Koehler, Phil Trinder, and Michel Steuwer. Sketch-guided equality saturation: Scaling equality saturation to complex optimizations in languages with bindings, 2021.

# 5 Working Groups

## 5.1 TTV in Tensorly

*Jeremy Cohen (CREATIS, CNRS, Villeurbanne, FR, jeremy.cohen@cnrs.fr)*
*Cem Bassoy (Fraunhofer IOSB, Ettlingen, DE)*
*Lawrence Mitchell (NVIDIA Corporation, Santa Clara, US, lmitchell@nvidia.com)*



Tensorly is a python library that provides high-level API for tensor decompositions [3]. It contains highly-efficient implementations of state-of-the-art algorithms for tensor decompositions. One key operation for the tensor decomposition is the so-called TTV (tensor-times-vector) operation. Tensorly's TTV implementation unfolds tensors in order to use efficiently implemented matrix-vector operations with unfolded tensors. The unfolding operation however consumes more memory than a naive implementation and introduces additional runtime overhead compared to high-performance tensor multiplication algorithms.

The goal of this breakout group was to investigate the usability of an high-performance open-source C++ implementation of TTV [1] and to estimate potential performance gains for Tensorly. One major benefit of Cem's C++ library is it's ability to provide fast tensor-vector multiplication for tensors that are stored according to last-order storage format which is (if not otherwise specified) the common format used in NumPy. We were able to integrate Cem's TTV library into Tensorly without modifying the original C++ code. We observed considerable speedups for certain tensor shapes. With promising results, we intend to continue our collaboration and to continuously report our latest findings in [2].

**References**

1 Cem Bassoy. Design of a high-performance tensor-vector multiplication with blas. In João M. F. Rodrigues, Pedro J. S. Cardoso, Jânio Monteiro, Roberto Lam, Valeria V. Krzhizhanovskaya, Michael H. Lees, Jack J. Dongarra, and Peter M.A. Sloot, editors, *Computational Science – ICCS 2019*, pages 32–45, Cham, 2019. Springer International Publishing.

2 `https://github.com/cohenjer/tensorly/tree/tensordot_and_ttv/ttv_and_tensordot`.

3 Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. Tensorly: Tensor learning in python. *Journal of Machine Learning Research (JMLR)*, 20(26), 2019.

## 5.2 A unified domain specific language for tensor contractions: Dagstuhl Tensor Language

*Albert Cohen (Google – Paris, FR, albertcohen@google.com)*
*Teodoro F. Collin (MIT CSAIL, US, teoc@mit.edu)*
*David Ham (Imperial College London, GB, david.ham@imperial.ac.uk)*
*Paul H. J. Kelly (Imperial College London, GB, p.kelly@imperial.ac.uk)*
*Thomas Koehler (University of Glasgow, GB, t.koehler.1@research.gla.ac.uk)*
*Norman A. Rink (DeepMind, London, GB, nrink@deepmind.com)*
*Edward Stow (Imperial College London, GB, edward.stow16@imperial.ac.uk)*
*Carsten Uphoff (Intel Deutschland GmbH, Feldkirchen, DE, carsten.uphoff@intel.com)*
*Sophia Vorderwuelbecke (Imperial College London, GB, s.vorderwuelbecke18@imperial.ac.uk)*
*Connor Ward (Imperial College London, GB, c.ward20@imperial.ac.uk)*



Dagstuhl Tensor Language (DTL) is the continuation of work done in the previous Dagstuhl Seminar, the aim being to bring together a common language at a highly abstract level that sufficiently describes the mathematics of tensor computations. The language is designed to be minimal and general in that it makes no attempt to specialise towards any hardware or algorithmic feature beyond tensor contractions; but with a view that is must become extensible and through directives or annotations a policy for decision making and optimisation for specific use cases could be produced.

### 5.2.1 Design

Our primary goal in the design of DTL was to build a minimal and unambiguous tensor language that targeted a minimal set of optimizations that could be done with just formulas for the sizes of the tensors. Strength reduction was the primary motivator. We were able to eventually sketch out a design by using an insight from several other languages for tensor computations: the need for an explicit unbind operator (an operator that builds a tensor by iterating over a given set of indexes) to avoid ambiguities in index notation. Several other projects, targeted at different problems (e.g Chiw's EIN for Diderot, GEM from Firedrake, Dex for ML, ATL for optimization problems) have postulated and to various extents used such an operation. We have incorporated it as a primary and fundamental feature for our minimal set of operations.

### 5.2.2 Open Problems

There are disagreements on the precise way we would like others to use this language. For some, we would hope this to basically be an IR and to be used in other people's stack after a point of lowering. For others, we want this to be a base language that is easily extended to other languages (i.e to be functorial in the sense of standard ML modules). In either case, we did not figure out how this language could easily modified or extended by users towards other means. In particular, although the base language works, we left several design issues unsolved that could be rephrased as areas where we need to enable extensibility via some means:

- Meta data associated to tensors and how it could be use by general strength reduction algorithms or passed down a compiler stack or used to call special math operations

- Special constant tensors with special algebraic relations (e.g. $\epsilon_{ijk...}$ and $\delta_{ij}$) or even languages for constant tensors built in some other expression language
- Arithmetic operations on indices, affine, modular, or maybe non-linear
- Special operations on tensor indices to transform them so that operations such as $Ax$ recreate convolution. For example, circular indexing and the topelitz transform employed in ML to do convolution with GEMM
- Non-linear mathematical operations on tensors as opposed to scalars (e.g inverse of a matrix)
  - Even if there should be linear operations on tensors
- Support for non-rectangular iteration spaces.
- Support for non-affine or even sparse iteration spaces.
- General support for iteration spaces
- Use of index expressions to make constant tensors (the algebraic use of indexes, allowing `i` to be promoted to a tensor $[0, 1, 2, 3, ..., n]$)
- Support for algebraic sparsity via conditionals in constant tensors and/or via more complex indexing operations
- Support for sequence spaces or semi-rings or other basic algebraic operations

We could probably generate a longer list, but the critical point is that the basic objects of the language (tensors, constants, operations, indexes, iteration spaces?) could be plausible extended in a variety of potentially overlapping ways (e.g. two of these extensions enable convolution support) or at least users might want to pass information about this stuff through a compilation pipeline that uses DTL (e.g. tagging where a tensor or operation comes from). We did not resolve these issues, but we recognize them and hope to solve them.

### 5.2.3 Future Work

One potential avenue we want to investigate is building this system within MLIR and seeing to what extent MLIR helps us. But even there, we must return to a central set of optimizations that we want to share, and we must also do the algorithmic work of how those optimizations could be extended if we allow the language to be extended along some of these axises. This was one of the core arguments against enabling too much extensionality as it makes the core rationale less reasonable.

### 5.2.4 Ongoing Work

Since the seminar, work has continued in rounding out a python-embedded implementation of the language with a project architecture that focuses on keeping the language minimal while allowing different back-ends to produce implementations or eagerly compute results. We have produced a work-in-progress reference implementation that uses python with Numpy for un-optimised but semantically correct outputs. In parallel, we have begun mapping the python-embedded implementation into an xDSL dialect with the intention to make use of existing work to enable lower levels of optimisations through xDSL's own linear algebra dialects and also further mapping into the MLIR compiler infrastructure.

## Participants


- Cem Bassoy
Fraunhofer IOSB – Ettlingen, DE
- Paolo Bientinesi
University of Umeå, SE
- Simon Bonér
University of Umeå, SE
- Albert Choen
Google – Paris, FR
- Jeremy Cohen
CNRS – IRISA – Rennes, FR
- Teodoro F. Collin
MIT – Cambridge, US
- Jutho Haegerman
Ghent University, BE
- David Ham
Imperial College London, GB
- Paul H. J. Kelly
Imperial College London, GB
- Thomas Koehler
University of Glasgow, GB
- Lawrance Mitchell
NVIDIA – Santa Clara, US
- Christos Psarras
RWTH Aachen, DE
- Norman A. Rink
Google DeepMind – London, GB
- P. Sadayappan
University of Utah, US
- Paul Springer
Nvidia Corp. – Santa Clara, US
- Edward Stow
Imperial College London, GB
- Volker Tresp
LMU München & Siemens, Müchen, DE
- Bora Ucar
CNRS and ENS Lyon, FR
- Carsten Uphoff
Intel Deutschland GmbH – Feldkirchen, DE
- Edward Valeev
Virginia Tech – Blacksburg, US
- Sophia Vorderwuelbecke
Imperial College London, GB
- Connor Ward
Imperial College London, GB


## Remote Participants


- Charisee Chiw
Google – San Francisco, US
- Nadav Cohen
Tel Aviv University, IL
- Edoardo Di Napoli
Jülich Supercomputing Centre, DE
- Rong Ge
Duke University – Durham, US
- Johnnie Gray
Caltech – Pasadena, US
- Vinod Grover
NVIDIA – Redmond, US
- Furong Huang
University of Maryland – College Park, US
- Katharina Kormann
Uppsala University, SE
- Jean Kossaifi
NVIDIA – Redmond, US
- Jiajia Li
Pacific Northwest National Lab. – Richland, US
- Devin Matthews
SMU – Dallas, US
- Luke Panayi
Imperial College London, GB
- Vivek Srikumar
University of Utah – Salt Lake City, US
- Edwin Miles Stoudenmire
Flatiron Institute – New York, US
- Richard M. Veras
University of Oklahoma – Norman, US
- Qi (Rose) Yu
University of California – San Diego, US
- Pan Zhang
Chinese Academy of Science – Beijing, CN

Report from Dagstuhl Seminar 22102

# Computational Models of Human-Automated Vehicle Interaction

**Christian P. Janssen*[1], Martin Baumann*[2], Antti Oulasvirta*[3], Shamsi Tamara Iqbal*[4], and Luisa Heinrich†[5]**

1 **Utrecht University, NL.** c.p.janssen@uu.nl
2 **Universität Ulm, DE.** martin.baumann@uni-ulm.de
3 **Aalto University, FI.** antti.oulasvirta@aalto.fi
4 **Microsoft – Redmond, US.** shamsi@microsoft.com
5 **Universität Ulm, DE.** luisa.heinrich@uni-ulm.de

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 22102 "Computational Models of Human-Automated Vehicle Interaction". At this Dagstuhl Seminar, we discussed how computational (cognitive) models can be used to model human-automated vehicle interaction. The seminar is motivated by developments in the field of semi-automated driving where humans and vehicles interact as teams to either both contribute to the drive (partnership) or to have safe transitions of control from vehicle to human and vice-versa. Computational (cognitive) models can be used in these situations to simulate or model human behavior and thought. Such models can be used among others to better understand human behavior, to test "what if" scenarios to guide design, or to even provide input to the vehicle about the human's potential behavior and thoughts. The seminar was attended by experts in various fields including computer science, cognitive science, engineering, automotive UI, human-computer interaction, and human factors. They represented academia, industry, and government organizations. With the attendees, we discussed five challenges of the field during panel discussion sessions:

- Challenge 1: How can models inform design and governmental policy?
- Challenge 2: What phenomena and driving scenarios need to be captured?
- Challenge 3: What technical capabilities do computational models possess?
- Challenge 4: How can models benefit from advances in AI while avoiding pitfalls?
- Challenge 5: What insights are needed for and from empirical research?

The attendees then split off into smaller working groups to discuss aspects of these challenges in more depth. Based on these discussions and other input from the attendees, this Dagstuhl report reports the following:

- an executive summary of the seminar
- position perspectives of all the attendees (section: "Talks")
- summaries of the various working groups (section: "Working Groups")
- summaries of the five panels (section: "Panel Discussions")
- an overview of relevant papers (section: "Open Problems")
- a research agenda with some of the most important developments and needs we identified for the field (section: "Open Problems")

All in all, we believe the seminar has shown that this field has lots of potential for development and an active community to tackle pressing issues. We can't wait to see what results the participants of the seminar will bring to the field in the future.

**Seminar** March 6–11, 2022 – http://www.dagstuhl.de/22102

* Editor / Organizer
† Editorial Assistant / Collector


Computational Models of Human-Automated Vehicle Interaction, *Dagstuhl Reports*, Vol. 12, Issue 3, pp. 15–81
Editors: Christian P. Janssen, Martin Baumann, Antti Oulasvirta, Shamsi Tamara Iqbal, and Luisa Heinrich
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

**2012 ACM Subject Classification** Human-centered computing → HCI design and evaluation methods; Human-centered computing → HCI theory, concepts and models; Human-centered computing → Human computer interaction (HCI); Human-centered computing; Human-centered computing → Interactive systems and tools; Human-centered computing → User models
**Keywords and phrases** artificial intelligence, automated vehicles, cognitive science, computational models, human-automation interaction, human-computer interaction, semi-automated vehicles, user models
**Digital Object Identifier** 10.4230/DagRep.12.3.15

# 1 Executive Summary

*Christian P. Janssen (Utrecht University, NL)*
*Martin Baumann (Universität Ulm, DE)*
*Shamsi Tamara Iqbal (Microsoft – Redmond, US)*
*Antti Oulasvirta (Aalto University, FI)*



This is the executive summary of Dagstuhl 22102: Computational Models of Human-Automated Vehicle Interaction, which took place March 6-11th 2022 in Hybrid format. The executive summary first summarizes the motivation of the seminar, then gives an overview of the broad challenges that were discussed, it then presents the results of the seminar. As this is only the summary, there are a lot more details about every item and result in other parts of this report, these are therefore referred to.

It has been a fruitful meeting, which sparked many research ideas. We want to thank all the attendees for their attendance and all the input they generate. We hope that it is of value to the community, and we can't wait to see what other results follow in the future based on discussions that started at this seminar!

Christian Janssen, Martin Baumann, Antti Oulasvirta, and Shamsi Iqbal (organizers)

## Computational Models of Human-Automated Vehicle Interaction: Summary of the field

The capabilities of automated vehicles are rapidly increasing, and are changing human interaction considerably (e.g., [4, 6, 29]). Despite this technological progress, the path to fully self-driving vehicles without any human intervention is long, and for the foreseeable future human interaction is still needed with automated vehicles (e.g., [15, 22, 29, 37, 48, 47]). The principles of human-automation interaction also guide the future outlook of the European Commission [13, 14]. Human-automated vehicle interaction can take at least two forms. One form is a partnership, in which the human and the automated vehicle both contribute in parallel to the control of the vehicle. Another form is in transitions of control, where the automated system at times takes over full control of the vehicle, but transitions control back to the human when desired by the human, or when required due to system limitations. For both the partnership and the transition paradigm it is beneficial when the car and the human have a good model of each other's capabilities and limitations. Accurate models can make clear how tasks are distributed between the human and the machine. This helps avoid misunderstandings, or mode confusion [45], and thereby reduces the likelihood of accidents and incidents. A key tool in this regard is the use of computational (cognitive) models: computational instantiations that simulate the human thought process and/or their interaction with an automated vehicle. Computational models build on a long tradition in cognitive science (e.g., [35, 36, 44]), human factors and human-computer interaction (e.g., [10, 39, 27], neuroscience (e.g., [12, 31]), and AI and engineering (e.g., [17, 42]). By now, there are a wide set of varieties that can be applied to different domains, ranging from constrained theoretical problems to capturing real-world interaction [38]. Computational models have many benefits. They enforce a working ethic of "understanding by building" and require precision in specification ([34], see also [8, 32, 41]). Models can test the impact of changes in parameters and assumptions, which allows for wider applicability and scalability (e.g., [2, 16, 44]). More generally, this allows for testing "what if" scenarios. For human-automated vehicle interaction in particular, it allows testing of future adaptive systems that are not yet on the road. Automated driving is a domain where computational models can be applied. Three approaches have only started to scratch the surface. First, the large majority of models focus on engineering aspects (e.g., computer vision, sensing the environment, flow of traffic) that do not consider the human extensively (e.g., [7, 18, 33]). Second, models that focus on the human mostly capture manual, non-automated driving (e.g., [44, 9, 25]). Third, models about human interaction in automated vehicles are either conceptual (e.g. [20, 22]) or qualitative, and do not benefit from the full set of advantages that computational models offer. In summary, there is a disconnect between the power and capabilities that computational models offer for the domain of automated driving, and today's state-of-the-art research. This is due to a set of broad challenges that the field is facing and that need to be tackled over the next 3-10 years, which we will discuss next.

## Description of the seminar topics and structure of the seminar report

The seminar topics were clustered around five broad challenges, for which we provide a brief description and example issues that were discussed addressed. Although the challenges are presented separately, they are interconnected and were discussed in an integrated manner during the seminar. During the seminar, each challenge was discussed in a panel, with all attendees taking part in at least one panel. After each panel, the group was split up in smaller workgroups, and discussed the themes in more lengths. The summary of each panel discussion can be found later in this report under the section "panel discussions". The outcomes of the workgroups can be found later in this report under the section "workgroups". In addition, all attendees wrote short abstracts that summarized their individual position.

### Challenge 1: How can models inform design and governmental policy?

Models are most useful if they are more than abstract, theoretical vehicles. They should not live in a vacuum, but be related to problems and issues in the real world. Therefore, we want to explicitly discuss how models can inform the design of (in-)vehicle technology, and how they can inform policy. As both of these topics can fill an entire Dagstuhl by themselves, our primary objective is to identify the most pressing issues and opportunities. For example, looking at:

- *Types of questions:* what types of questions exist at a design and policy level about human-automated vehicle interaction?
- *How to inform decisions:* How can models be used to inform design and policy decisions? What level of detail is needed here? What are examples of good practices?
- *Integration:* Integration can be considered in multiple ways. First, how can ideas from different disciplines be integrated (e.g., behavioral sciences, engineering, economics), even if they have at times opposing views (e.g., monetary gains versus accuracy and rigor)? Second, how can models become better integrated in the design and development process as tools to evaluate prototypes (instead of running empirical tests)? And third, how can models be integrated into the automation (e.g., as a user model) to broaden the automation functionality (e.g., prediction of possible driver actions, time needed to take over)?

### Challenge 2: What phenomena and driving scenarios need to be captured?

The aim here is to both advance theory on human-automation interaction while also contributing to understanding realistic case studies for human-automation interaction that are faced for example by industry and governments. The following are example phenomena:

- *Transitions of control and dynamic attention:* When semi-automated vehicles transition control of the car back to the human, they require accurate estimates of a user's attention level and capability to take control (e.g., [22, 49]).
- *Mental models, machine models, mode confusion, and training and skill:* Models can be used to estimate human's understanding of the machine and vice-versa (e.g., [20]). Similarly, they might be used to estimate a human driver's skill level, and whether training is desired.
- *Shared control:* In all these scenarios, there is some form of shared control. Shared control requires a mutual understanding of human and automation. Computational models can be used to provide such understanding for the automation (e.g., [50]).

### Challenge 3: What technical capabilities do computational models possess?

A second challenge has to do with the technical capabilities of the models. Although the nature of different modeling frameworks and different studies might differ [38], what do we consider the core functionality? For example, related to:

- *Compatibility:* To what degree do models need to be compatible with simulator software (e.g., to test a "virtual participant"), hardware (e.g., be able to drive a car on a test track), and other models of human thinking?
- *Adaptive nature:* Computational models aim to strike a balance between precise predictions for more static environments and being able to handle open-ended dynamic environments (like everyday traffic). How can precision be guaranteed in static and dynamic environments? How can models adapt to changing circumstances?
- *Speed of development and broader adoption:* The development of computational models requires expertise and time. How can development speed be improved? How can communities benefit from each other's expertise?

### Challenge 4: How can models benefit from advances in AI while avoiding pitfalls?

At the moment there are many developments in AI that computational models can benefit from. Three examples are advances in (1) simulator-based inference (e.g., [26]) to reason about possible future worlds (e.g., varieties of traffic environments), (2) reinforcement learning [46] and its application to robotics [30] and human driving [25], and (3) deep learning [17] and its potential to predict driver state or behavior from sensor data. At the same time, incorporation of AI techniques also comes with challenges that need to be addressed. For example:

- *Explainability:* Machine learning techniques are good at classifying data, but do not always provide insight into why classifications are made. This limits their explainability and is at odds with the objective of computational models to gain insight into human behavior. How can algorithms' explainability be improved?
- *Scalability and generalization*: How can models be made that are scalable to other domains and that are not overtrained on specific instances? How can they account for future scenarios where human behavior might be hard to predict [5]?
- *System training and corrective feedback:* if models are trained on a dataset, what is the right level of feedback to correct an incorrect action to the model? How can important new instances and examples be given more weight to update the model's understanding without biasing the impact?

### Challenge 5: What insights are needed for and from empirical research?

Models are only as good to the degree as they can describe and predict phenomena in the real world. Therefore, empirical tests are an important consideration. Example considerations are:

- *Capturing behavioral change and long-term phenomena:* Many current computational models capture the results of a single experiment. However, behavior might change with more exposure to and experience with automated technology. How can such (long-term) behavior change be tested?
- *Capturing unknown future scenarios:* Many automated technologies that might benefit from computational models are not yet commercially available. How can these best be studied and connected to computational models?
- *Simulated driving versus real-world encounters:* To what degree are simulator tests representative of real-world scenarios (e.g., [19])?

## Results

The seminar has generated the following results.

1. **Overview of state-of-the-art technologies, methods, and models.** The spectrum of computational modeling techniques is large [38, 21, 24]. Before and during the conference, we have discussed various methods and techniques. In particular, this report contains a dedicated chapter called "Relevant papers for modeling human-automated vehicle interaction" in which we report a long set of papers that the community identified as being relevant to the field. We encourage scholars to take a look at it.
2. **List of grand challenges with solution paths.** We have identified five grand challenges and discussed those in detail during the panels. Our chapters on "panel discussions" report the outcomes of these discussions. Moreover, the workgroups further report the in-depth discussions that smaller groups had about these challenges. See the section "working group" of this report. The results only start to scratch the surface of some of the grand challenges for the application of computational cognitive modeling that need to

be faced within the next 3 to 10 years, and their paths to solutions. Based on discussions, groups of authors plan to work on more papers and workshops around topics that they deemed worthy of further discussion. For example, we discussed whether there are specific driving scenarios that a computational model should be able to capture, and how success might be quantified (e.g., whether these challenges should take the form of competitions, akin to DARPA's Grand Challenge for automated vehicles [11] or "Newell's test" for cognitive models [3]).

3. **Research agenda to further the field.** This report also reports a research agenda that is intended to further the field. For each specific grand challenge, we have identified more specific areas of research that need futher exploration. We refer to the dedicated section in this report called "Research agenda to further the field". The organizers of the seminar will also organize a dedicated journal special issue around the topic, in which further results that arose from the seminar can be reported.

### References


1 Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Iqbal, S.T., and Teevan, J. (2019). Guidelines for human-AI interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1-13).

2 Anderson, J. R. (2007). How can the human mind occur in the physical universe? (Vol. 3). Oxford University Press.

3 Anderson, J. R., and Lebiere, C. (2003). The Newell test for a theory of cognition. Behavioral and brain Sciences, 26(5), 587-601.

4 Ayoub, J., Zhou, F., Bao, S., and Yang, X. J. (2019). From Manual Driving to Automated Driving: A Review of 10 Years of AutoUI. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 70-90).

5 Bainbridge, L. (1983). Ironies of automation. In Analysis, design and evaluation of man–machine systems (pp. 129-135). Pergamon.

6 Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C., and Winner, H. (2014). Three decades of driver assistance systems: Review and future perspectives. IEEE Intelligent Transportation Systems Magazine, 6(4), 6–22.

7 Brackstone, M., and McDonald, M. (1999). Car-following: a historical review. Transportation Research Part F: Traffic Psychology and Behaviour, 2(4), 181-196.

8 Brooks, R. A. (1991). Intelligence without representation. Artificial intelligence, 47(1-3), 139-159.

9 Brumby, D. P., Janssen, C. P., Kujala, T., and Salvucci, D. D. (2018). Computational models of user multitasking. Computational interaction design, 341-362.

10 Card, S. K., Moran, T., and Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: L. Erlbaum Associates Inc.

11 Darpa (2020) The Grand Challenge. Accessed online on July 6 2020 at `https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles`

12 Eliasmith, C. (2013). How to build a brain: A neural architecture for biological cognition. Oxford University Press.

13 European Commission (2018, 17 May). On the road to automated mobility: An EU strategy for mobility of the future (pp. 1–17). Brussels, BE. Communication COM(2018) 283 final.

14 European Commission (2020, 19 February). Shaping Europe's digital future. Brussels (BE). Communication COM(2020) 67 final.

15 Favarò, F. M. (2020). Unsettled Issues Concerning Semi-Automated Vehicles: Safety and Human Interactions on the Road to Full Autonomy. Technical report for the SAE. Warrendale, PA: SAE International. Retrieved from https://doi.org/s://www.sae.org/publications/technical-papers/content/epr2020001/

**16** Gray, W. D. (Ed.). (2007). Integrated models of cognitive systems (Vol. 1). Oxford University Press.
**17** Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. Cambridge, MA: MIT press.
**18** Helbing, D. (2001). Traffic and related self-driven many-particle systems. Reviews of modern physics, 73(4), 1067.
**19** Hock, P., Kraus, J., Babel, F., Walch, M., Rukzio, E., and Baumann, M. (2018). How to design valid simulator studies for investigating user experience in automated driving – Review and hands-on considerations. Proceedings of the International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, 105–117. New York, NY: ACM Press
**20** Janssen, C. P., Boyle, L. N., Kun, A. L., Ju, W., and Chuang, L. L. (2019). A Hidden Markov Framework to Capture Human–Machine Interaction in Automated Vehicles. International Journal of Human-Computer Interaction, 35(11), 947–955.
**21** Janssen, C. P., Boyle, L. N., Ju, W., Riener, A., and Alvarez, I. (2020). Agents, environments, scenarios: A framework for examining models and simulations of human-vehicle interaction. Transportation research interdisciplinary perspectives, 8, 100214.
**22** Janssen, C. P., Iqbal, S. T., Kun, A. L., and Donker, S. F. (2019). Interrupted by my car? Implications of interruption and interleaving research for automated vehicles. International Journal of Human-Computer Studies, 130, 221–233.
**23** Janssen, C. P., and Kun, A. L. (2020). Automated driving: getting and keeping the human in the loop. Interactions, 27(2), 62-65.
**24** Jeon, M., Zhang, Y., Jeong, H., P. Janssen, C.P., and Bao, S. (2021). Computational Modeling of Driving Behaviors: Challenges and Approaches. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 160-163).
**25** Jokinen, J.P.P., Kujala, T., and Oulasvirta, A. (2021) Multitasking in Driving as Optimal Adaptation under Uncertainty. Human Factors, , 63(8), 1324-1341.
**26** Kangasrääsiö, A., Jokinen, J. P., Oulasvirta, A., Howes, A., and Kaski, S. (2019). Parameter inference for computational cognitive models with Approximate Bayesian Computation. Cognitive Science, 43(6), e12738.
**27** Kieras, D. (2012). Model-based evaluation. In: Jacko and Sears (Eds.) The Human-Computer Interaction Handbook (3rd edition), 1294-310. Taylor and Francis
**28** Kun, A. L. (2018). Human-Machine Interaction for Vehicles: Review and Outlook. Foundations and Trends in Human-Computer Interaction, 11(4), 201–293.
**29** Kun, A. L., Boll, S., and Schmidt, A. (2016). Shifting Gears: User Interfaces in the Age of Autonomous Vehicles. IEEE Pervasive Computing, 32–38.
**30** Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. arXiv preprint arXiv:1805.00909.
**31** Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco, CA: W.A. Freeman.
**32** McClelland, J. L. (2009). The place of modeling in cognitive science. Topics in Cognitive Science, 1(1), 11-38.
**33** Mogelmose, A., Trivedi, M. M., and Moeslund, T. B. (2012). Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. IEEE Transactions on Intelligent Transportation Systems, 13(4), 1484-1497.
**34** Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium.In Chase (ed.) Visual Information Processing. New York: Academic Press.
**35** Newell, A. (1990). Unified theories of cognition. Cambridge, MA: Harvard University Press.

36 Newell, A., and Simon, H. A. (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall
37 Noy, I. Y., Shinar, D., and Horrey, W. J. (2018). Automated driving: Safety blind spots. Safety science, 102, 68-78.
38 Oulasvirta, A. (2019). It's time to rediscover HCI models. Interactions, 26(4), 52-56.
39 Oulasvirta, A., Bi, X., Kristensson, P-O., and Howes, A., (Eds.) (2018). Computational Interaction. Oxford University Press
40 Peebles, D., and Cooper, R. P. (2015). Thirty years after Marr's vision: levels of analysis in cognitive science. Topics in cognitive science, 7(2), 187-190.
41 Pfeifer, R., and Scheier, C. (2001). Understanding intelligence. Cambridge, MA: MIT press.
42 Russell, S., and Norvig, P. (2002). Artificial intelligence: a modern approach. Uppersaddle River, NJ: Pearson
43 SAE International. (2014). J3016: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. Warrendale, PA, USA: SAE International
44 Salvucci, D. D., and Taatgen, N. A. (2011). The multitasking mind. Oxford University Press.
45 Sarter, N. B., and Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Human factors, 37(1), 5-19.
46 Sutton, R., and Barto, A. G. (2018). Reinforcement learning: An introduction. Cambridge, MA: MIT Press
47 Walch, M., Sieber, T., Hock, P., Baumann, M., and Weber, M. (2016). Towards cooperative driving: Involving the driver in an autonomous vehicle's decision making. In Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 261–268. New York, NY: ACM Press.
48 Walch, M., Mühl, K., Kraus, J., Stoll, T., Baumann, M., and Weber, M. (2017). From Car-Driver-Handovers to Cooperative Interfaces: Visions for Driver–Vehicle Interaction in Automated Driving. In G. Meixner and C. Müller (Eds.): Automotive User Interfaces: Creating Interactive Experiences in the Car (pp. 273–294). Springer International Publishing.
49 Wintersberger, P., Schartmüller, C., and Riener, A. (2019). Attentive User Interfaces to Improve Multitasking and Take-Over Performance in Automated Driving: The Auto-Net of Things. International Journal of Mobile Human Computer Interaction, 11(3), 40-58.
50 Yan, F., Eilers, M., Weber, L., and Baumann, M. (2019). Investigating Initial Driver Intention on Overtaking on Rural Roads. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (pp. 4354-4359). IEEE.

# 2 Table of Contents

**Executive Summary**
*Christian P. Janssen, Martin Baumann, Shamsi Tamara Iqbal, and Antti Oulasvirta* 16

**Overview of Talks**

Computational modeling is key for advancing knowledge about human-AV interaction
*Martin Baumann* . . . 27

Automated Driving and Cognitive Architecture Models
*Jelmer Borst* . . . 27

Reflections on 15 years of Modelling In-Car Multitasking Behaviour
*Duncan Brumby* . . . 28

Modeling for AV-RU interaction
*Debargha Dey* . . . 29

Data cautions in the use of machine learning tools
*Birsen Donmez* . . . 30

Triangulation of Cognitive Model and ML-based models
*Patrick Ebel* . . . 30

Insights from Language Modelling Research
*Justin Edwards* . . . 31

Towards a holistic architecture of human driver behavior
*Mark Eilers* . . . 31

Reliable Autonomy Based on Imperfect Models
*Martin Fränzle* . . . 32

How to model situation awareness and predict driver take-over ability
*Luisa Heinrich* . . . 33

Hybrid models – integrating Machine learning approaches into cognitive models
*Moritz Held* . . . 33

The Nomadic Worker: Autonomous Vehicles as Future Worksites
*Shamsi Tamara Iqbal* . . . 34

An exciting time for the field of Cognitive Modeling of Human-Automated Vehicle Interaction
*Christian P. Janssen* . . . 35

Technical capabilities for computational models
*Myounghoon Jeon* . . . 36

HMI design for autonomous vehicles: methodologies and intercultural analyses
*Xiaobei Jiang* . . . 36

Benefits and Challenges of Computational Cognitive Models
*Jussi Jokinen* . . . 37

Much more data is needed
*Wendy Ju* . . . 37

Computational cognitive models for attention monitoring
*Tuomo Kujala* . . . 38

In-vehicle human-automation interaction: Bumper-to-bumper traffic, and short bursts of activity in non-driving tasks
*Andrew Kun* . . . 39

Four challenges to computational models in Human-AV interaction
*Dietrich Manstetten* . . . 39

Computational models of humans as tools for enabling safe and acceptable vehicle automation
*Gustav Markkula* . . . 40

Data for and models on Trained Operators
*Nikolas Martelaro* . . . 40

Simulation Intelligence and Computational Models in Autonomous Vehicles
*Roderick Murray-Smith* . . . 41

Computational Rationality as an Emerging Approach to Inform the Design of Interactive System
*Antti Oulasvirta* . . . 41

A common understanding of models...
*Andreas Riener* . . . 42

Anticipating the cognitive state of the driver
*Nele Rußwinkel* . . . 43

Models that Inform Design
*Shadan Sadeghian Borojeni* . . . 43

Complexity of computational models
*Hatice Sahin* . . . 44

Teamwork and Driver-Vehicle Cooperation
*Boris van Waterschoot* . . . 44

From Trust Studies to Trust Models
*Philipp Wintersberger* . . . 45

Understanding and challenges of computional models of human-automated vehicles interaction
*Fei Yan* . . . 46

**Working groups**

Learning and Adaptation
*Jelmer Borst, Alexandra Bremers, Birsen Donmez, Mark Eilers, and Roderick Murray-Smith* . . . 46

How can model use be increased in design?
*Alexandra Bremers, Jussi Jokinen, and Gustav Markkula* . . . 47

How can empirical research support computational models (and vice-versa)?
*Debargha Dey, Moritz Held, and Fei Yan* . . . 47

Modeling Long-term Effects in Human Technology Engagement
*Patrick Ebel, Hatice Sahin, and Philipp Wintersberger* . . . 48

Focused session "Application in specific scenarios"
*Martin Fränzle, Luisa Heinrich, and Roderick Murray-Smith* . . . . . 49

Towards common tests for Models
*Christian P. Janssen, Jelmer Borst, Benjamin Cowan, Justin Edwards, Mark Eilers, and Jussi Jokinen* . . . . . 50

What are important scenarios?
*Dietrich Manstetten, Moritz Held, Nele Rußwinkel, and Fei Yan* . . . . . 51

What is needed for different SAE level?
*Dietrich Manstetten, Martin Baumann, and Shadan Sadeghian Borojeni* . . . . . 52

Multi-agent modelling: Platooning use case
*Gustav Markkula, Lewis Chuang, Debargha Dey, Patrick Ebel, and Christian P. Janssen* . . . . . 52

What computational models are available for human-automation interaction?
*Gustav Markkula, Debargha Dey, Moritz Held, Nele Rußwinkel, and Fei Yan* . . . . . 55

Modeling workflows
*Antti Oulasvirta, Alexandra Bremers, and Tuomo Kujala* . . . . . 55

What is prediction?
*Antti Oulasvirta, Jelmer Borst, Patrick Ebel, Martin Fränzle, and Nele Rußwinkel* 56

Which models should be used and what should be their content?
*Shadan Sadeghian Borojeni, Martin Baumann, Luisa Heinrich, Dietrich Manstetten, Roderick Murray-Smith, and Hatice Sahin* . . . . . 57

Which scenarios should be taken into account in computational models of human-automated vehicle interaction?
*Hatice Sahin, Duncan Brumby, Jussi Jokinen, and Shadan Sadeghian Borojeni* . . 58

Modeling Trust in Automation
*Philipp Wintersberger, Martin Baumann, Justin Edwards, Luisa Heinrich, and Tuomo Kujala* . . . . . 58

**Panel discussions**

Summary of Panel 1: How can models inform design?
*Antti Oulasvirta, Alexandra Bremers, Lewis Chuang, Debargha Dey, Andreas Riener, and Shadan Sadeghian Borojeni* . . . . . 59

Summary of Panel 2: What phenomena and driving scenarios need to be captured in computational models of human-automated vehicle interaction?
*Martin Baumann, Luisa Heinrich, Andrew Kun, Dietrich Manstetten, Nikolas Martelaro, and Hatice Sahin* . . . . . 60

Summary of Panel 3: What technical capabilities do computational models need to possess?
*Antti Oulasvirta, Jelmer Borst, Martin Fränzle, Myounghoon Jeon, Otto Lappi, Gustav Markkula, and Nele Rußwinkel* . . . . . 64

Summary of Panel 4: How can models benefit from advances in AI while avoiding its pitfalls?
*Christian P. Janssen, Duncan Brumby, Birsen Donmez, Justin Edwards, Mark Eilers, Moritz Held, Jussi Jokinen, and Roderick Murray-Smith* . . . . . 65

Summary of Panel 5: What insights are needed for or from empirical research?
*Shamsi Tamara Iqbal, Linda Ng Boyle, Benjamin Cowan, Patrick Ebel, Wendy Ju, Tuomo Kujala, Philipp Wintersberger, and Fei Yan* . . . . . . . . . . . . . . . . . . . 69

**Open problems**

Relevant papers for modeling human-automated vehicle interaction
*Christian P. Janssen* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 72

Research agenda to further the field
*Christian P. Janssen, Martin Baumann, Shamsi Tamara Iqbal, and Antti Oulasvirta* 77

**Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

**Remote Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

# 3 Overview of Talks

## 3.1 Computational modeling is key for advancing knowledge about human-AV interaction

*Martin Baumann (Universität Ulm, DE)*



The recent technological development is bringing the vision of automated driving within realistic reach. But for really exploiting the potential of this technology in terms of safety, efficiency and comfort human factors knowledge has to be considered and integrated into the design of these systems. Designing the interaction of humans both inside and outside with automated vehicles (AVs) is therefore a key factor for the technology's success. Unfortunately, we still lack important knowledge about the psychological mechanisms underlying the interaction of humans with AVs. At least in some cases, this is not due to the fact that we miss empirical data, but because we miss precise definitions and theories of the relevant mechanisms. This is where computational models of human-automated vehicle interaction come in and might play a key role in advancing the theoretical basis of our discipline in this field.

I see mainly three relevant phenomena that are essential for understanding human-AV interaction and that might profit from computational modelling:

**i)** how do humans construct and maintain an adequate comprehension of the current situation including system status,

**ii)** what are the long-term effects of interacting with AVs, and one aspect that is especially important here is how does trust into AVs evolve over time. Building computational models of trust and its development will definitely help us move forward regarding this supposed to be highly relevant but only loosely defined concept.

**iii)** how do humans interact and cooperate with each other in complex and realistic traffic scenarios. Current research results are mainly based on simple situations and one to one interaction scenarios. To tackle the complexity of such situations computational models allowing to simulate and to understand the processes in deeper level might be very helpful.

I hope that this Dagstuhl Seminar will bring us a step closer to solutions for these challenges.

## 3.2 Automated Driving and Cognitive Architecture Models

*Jelmer Borst (University of Groningen, NL)*



**Main reference** Dario D. Salvucci: "Modeling Driver Behavior in a Cognitive Architecture", Hum. Factors, Vol. 48(2), pp. 362–380, 2006.
**URL** https://doi.org/10.1518/001872006777724417

High-level cognitive architectures such as ACT-R can be used to simulate human driving under various circumstances [1]. Although they give a faithful characterization of the cognitive system, due to their rule-based nature they seem to be too brittle to operate an automated vehicle. Instead, I argue for four use cases:

1) understanding human cognition in driving;
2) a hybrid cognitive model-machine learning system, where the cognitive model informs the machine-learning part about expected behavior of other drivers and accompanying proper behavior;
3) a model-tracing approach, where the cognitive model is used to warn the driver when deviating from predicted behavior; and
4) a model-tracing approach where the cognitive model increases the automation level of the car when high workload is predicted for the driver.

**References**

1 Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors*, 48(2), 362–380. https://doi.org/10.1518/001872006777724417.

## 3.3 Reflections on 15 years of Modelling In-Car Multitasking Behaviour

*Duncan Brumby (University College London, GB)*



It was 15 years ago that I first wrote a computational cognitive model of driving. The model was used to make performance predictions for a range of different dual-task interleaving strategies that a driver could potentially adopt when entering a number to a mobile phone while driving. This work was published at CHI 2007 and it provided a great foundation for a lot of future research, including Chris Janssen's PhD thesis (who is now one of the organisers of this meeting).

At this meeting, I've been reflecting on three significant advances that have happened since 2007.

First, many modern cars have advanced driver support systems. This includes cameras that monitor the road so that the car itself can manage speed and lateral control. These systems have become quite wide spread, and show that the basic demands of the driving task are changing.

Second, mobile devices have got better. In 2007, the iPhone had only just been released. Now many cars have in-car displays to show route directions, and integrated voice user interfaces that can be used to play music or send messages. The secondary tasks that drivers can do are also changing.

Third, the basic science of modelling drivers has advanced. In 2007 we were just beginning to explore ways to use reinforcement learning methods to get our models to adapt strategies for driving and secondary task interactions. At the time we opted for a simple "black box" approach. Since this time the great AI Spring has delivered modern machine learning techniques that can be use by our models. The work being done by Jussi Jokinen and Antti Oulasvirta is now realising what we only hoped would have been possible 15 years ago. The models that we can develop are changing.

The aim of this meeting then is to reflect on the advances that we've seen in these three inter-related areas that are relevant for developing Computational Models of Human-Automated Vehicle Interaction: the driving task, the secondary task, and the basic science of cognitive modelling. I hope as a result of this meeting we can set out a research agenda for the next 15 years of change.

## 3.4 Modeling for AV-RU interaction

*Debargha Dey (TU Eindhoven, NL)*



Modeling an environment or human behavior, in essence, is looking at the world as a computer. But the world is complex, and within the dynamism we operate, it is often hard to quantify the specific dimensions of interest that provoke a certain behavior we observe. A complex world emerges from a complex interaction from an unquantifiable explosion of parameters, which tend to make any mathematical solution intractable. The job of a cognitive/ computation model is to simplify this quagmire of parameters and attempt to make some sense of this complexity with some informed guess, or in other words, illuminate the black box. Empirical research can help in systematically investigating the hierarchy of effect of the different parameters.

In the world of interactions within automotive human factors, we can say a lot about situations and negotiations, but we are not able to draw boundaries on problems and definitively say how to solve them. Furthermore, when we attempt to answer these questions through models, the expected reaction of the environment (e.g. driver, pedestrian, vehicle, etc.) are often not captured. Applying this to the context of my current research domain of eHMIs (external Human-Machine Interfaces) that facilitate AV-road user communication, linear models tell us how pedestrians will react and how a car should interact. However, the interaction from a car can cause further behavioral adaptations in a pedestrian, which current models are often unable to capture. The concept of interdependence – collaboration, coordination, and teamwork – is a grand challenge that seems to emerge as a gap in the state of the art.

Specifically in the context of eHMI research, the biggest potential benefit of a modeling approach emerges when tackling the problem of scalability. The state of the art has been mainly confined to a "one-car-one-pedestrian" setup in conducting empirical research. This is primarily because when investigating scalability of interactions, even at the minimum viable condition of testing scalability (i.e. just two pedestrians), the complexity introduced leads to the previously mentioned explosion of parameters, which make systematic empirical studies untenable. A potential approach is therefore to use a variety of theory- and naturalistic-data driven models to identify and decode the interactions in multi-agent scenarios, and conduct empirical studies with a more informed, limited set of parameters.

Another application for models is as a filtering tool. If a set of data does not fit a proven model derived from data or theory, it can be used as an indication of an outlier, which can be applied in practice as an alerting system. However, model interpretability is an important and non-trivial part of the equation if a model is to be used in that way: a machine-learned black-box model needs to be interrogated by a surrogate interpretable/ cognitive model. Furthermore, the question "how do we evaluate a model" stays wide open. What is the metric of a good model? And what is the appropriate thing to model for? The present? The future? A specific, imagined future? Points to ponder.

## 3.5 Data cautions in the use of machine learning tools

*Birsen Donmez (University of Toronto, CA)*



For some applications, like real-time driver state detection, the power of black-box Machine Learning (ML) tools is indisputable over other types of driver models. These ML tools enable the predictive modeling of complex nonlinear systems and can accommodate the modeling of a large number of predictor variables and their interactions. However, such complex models also require large datasets, which may or may not be directly available to researchers.

Caution should be taken when applying ML approaches to small datasets that empirical researchers traditionally collect. Although data may be collected at high frequency from each participant, experimental/observational scientific studies are generally limited in sample size (i.e., number of participants). Traditional statistical modeling techniques (model building and validation) were developed for small samples, whereas ML techniques assume large amounts of data. For example, researchers who have data from an experiment may have to split their training/test datasets within participants rather than across participants. But this approach can create data leakage resulting in ML models being rewarded for identifying participants rather than signals associated with the phenomenon of interest.

Caution should also be taken when researchers are lucky enough to obtain large datasets from other sources (e.g., OEMs). Real world data is messy, especially if there is not much control on how data is collected/selected for model training – leading up to examples such as facial recognition algorithms that are inequitable. Empirical researchers who design their own data collection are in a good position to apply their expertise and not fall into the same pitfalls that others do, by questioning how the data was collected, whether it is representative, and whether correlations exist in the data.

## 3.6 Triangulation of Cognitive Model and ML-based models

*Patrick Ebel (Universität Köln, DE)*



Out of the many extremely interesting topics that were discussed, what is of greatest interest to me is the question of how to combine cognitive models with learned (ML-based) models. How can we theoretical/mechanistic approaches improve machine learning-based approaches and vice versa? An effective combination of both could be the key to addressing the accuracy vs. Interpretability trade-off in HCI. In my opinion, an interesting way to go is to explore how complex interactions can be broken down into multiple subtasks/modules that only describe a small part of human behavior (glance allocation, pointing time...). These modules need to be restricted such that they only operate within certain mechanistic/cognitive boundaries. By doing so one can understand the different mechanisms of human behavior that in their combination describe a "complete" human-machine interaction without sacrificing accuracy due to abstract (but theoretically sound) models that lack prediction accuracy. However, even though this idea might be tempting, each module would still be trained in a specific context which leads to the problem of generalizability and raises the question of uncertainty prediction. I think that the effective usage of large natural data can be a solution to this problem. Leveraging large amounts of data we can represent different driving situations and contexts that would lead to an increase in generalizability.

## 3.7 Insights from Language Modelling Research

*Justin Edwards (ADAPT Centre – Dublin, IE)*



For decades, language models were rule-based and theory-driven. Language data is and has been plentifully accessible – models were not theory-driven in the 1960s because in-the-wild language data was not available, but because we didn't have the compute power to make use of large amounts of language data. The last decade has been very different for language models, with very large models like BERT and GPT-3 which, by utilizing tremendous power to train the models, modelling language in a quite complex, data-driven way. These data driven models have surpassed theory-driven models in many language tasks. They have been deeply flawed in other ways however, because of attempts to generalize data-driven models when the sample of language data is not general or ought not be generalized. This has resulted in applications, like chatbots that use racist language, or applications said to be capable of making moral judgments despite much of the training data coming from a Reddit advice forum. As new types of in-the-wild human behavioural data becomes available for contexts like human behaviour in SAE Levels 3 and 4 of driving, it is crucial that we examine the biases inherent to our data sources – the things that either will not or ought not generalize – and to remember that these biases can be caried through in our models. We must choose model architectures which are sensitive to biases in are data and we must be careful in how generally we try to apply models trained on idiosyncratic human behavioural data.

## 3.8 Towards a holistic architecture of human driver behavior

*Mark Eilers (Humatects – Oldenburg, DE)*



Intelligent design of human-automated vehicle interaction requires computational models of the human driving behavior and cognition. Particularly holistic models that can explain and predict multiple aspects of SAE Level 0 human driving behavior simultaneously seem equally important as modelling tool during the design of driving automation systems and their certification, and as a component of the system itself. Unfortunately, there is still a lack in commonly agreed upon models or architectures that can bear such a title.

As extensively discussed during the seminar, there is a common request for open-access and agreed-upon datasets, scenarios, benchmarks, and competitions, and first plans emerged to address this issue in the future. I'd argue that this endeavor should be accompanied by attempts to define (and provide software support for) an open-access architecture for a holistic model of human driving behavior that could evolve into a kind of gold-standard of driver models.

For a recent starting point, [1] provide a framework for a unified visuomotor model of a driver's lateral control during simulated driving that couples gaze and steering control in a three-layered architecture. It would be interesting to see this framework extended and harmonized with theories for e.g., longitudinal control, intention formation, route planning, and situational awareness, to name a few.

Such an architecture could start as a collection of (potentially competing) models that implement different aspects of human driving behavior (e.g., [2] two-point visual control model of steering) that could be plugged together and interfaced with the currently envisioned datasets, scenarios, benchmarks, and challenges in open-access driving simulators.

**References**

1 Lappi, O. and Mole, C. D. (2018). Visuomotor control, eye movements, and steering: A unified approach for incorporating feedback, feedforward, and internal models. *Psychological Bulletin*,144(10), pp. 981-1001. https://doi.org/10.1037/bul0000150

2 Salvucci, D. D. and Gray, R. (2004). A two-point visual control model of steering. *Perception*, 33 (10), pp. 1233-1248.

## 3.9 Reliable Autonomy Based on Imperfect Models

*Martin Fränzle (Universität Oldenburg, DE)*



From the perspective of design and safety analysis of highly critical human-cyber-physical systems, like automated driving at SAE levels 4 and 5, the demands on epistemic validity and accuracy of human models running as human state predictors in the car might seem prohibitive and the search for appropriate human models elusive. In my contribution to the panel, I made the point that this may not actually be true: if the models generate predictions which permit reliable (yet probably pessimistic) inferences in case of uncertainty then these models can seamlessly be embedded into safety-oriented system architectures when resorting to techniques of safe planning and control under uncertainty. Control theory as well as formal methods in computer science have in the past exposed various fundamental approaches to rigorous handling of uncertainties, which the human models and their corresponding execution mechanisms would then, however, need to be able to interface to.

These approaches rely on comprehensive mathematical representation of uncertainties concerning system state and evolution, building on possible world semantics in that they represent a system by sets or distributions of possible states at any point of current and extrapolated future time. This may seem a minor variation from the currently prevalent state-based models, given that the latter frequently come in stochastic variants which implicitly define the required distributions. The interfacing problem, however, results from the execution mechanisms employed for computing state extrapolations over these models: simulation-oriented mechanisms generating state traces, even if they do so using means of randomized simulation for stochastic models, are not suitable, as approximating a distribution via massively iterated randomized simulation is grossly inadequate for embedding in hard real-time into, as necessary for an online mechanism. What we instead need are mechanisms directly computing distributions in a time-resolved manner.

We argue that using such mechanisms, quite unreliable or, more precisely, uncertain human models -in experiments we have used such with just 61% predictive accuracy- could effectuate significant safety gains in embedded control applications. Technically, the human model here acts inside the embedded control as an online proxy of the human. This proxy enables predictive evaluation of the consequences of possible control actions, thus permitting the technical system to pursue rolling-horizon model-predictive control that is provident to the human.

Such an online embedding of human models would, however, require that the models to be embedded feature rigorous real-time guarantees, that they can incorporate the (uncertain) state evidence provided by in-situ measurements, that they can exploit the latter for computing best-possible estimates of current state, and that the provide time-resolved distributional state predictions over reasonable horizons of the imminent future. All these are prerequisites for interfacing seamlessly to rolling-horizon model-predictive control. Current models or rather their execution mechanisms fail to meet these by primarily being targeted to simulation, i.e., to trace generation rather than to state estimation by distributions.

## 3.10 How to model situation awareness and predict driver take-over ability

*Luisa Heinrich (Universität Ulm, DE)*



With higher levels of automation, the human's role shifts from actively steering the vehicle to passively monitoring the system. However, as long as fully automated driving (SAE L5) is not available, humans will have to intervene in the driving task at times. Scenarios in which transitions are likely to occur should therefore receive special attention. For example, when exiting a highway, or when entering urban traffic. Starting with SAE L3 of automation, the human driver may disengage from the driving task and perform other non-driving related activities. When a TOR is issued, he or she may be out-of-the-loop and incapable of understanding the situation and acting appropriately to traffic events, especially in safety-critical situations with small time-budgets. In order to support the human during transitions in automated driving, we need to predict the driver state. Does the human have enough situation awareness to safely take over control of the vehicle? With situation awareness, I don't mean being able to consciously reproduce knowledge of the current situation, but having a well enough understanding to make appropriate driving decisions – implicit rather than explicit knowledge. Computational models may inform us about driver state by considering metrics relevant to the build-up of situation awareness, for example eye-tracking (behavioral) or physiological data (e.g., heart rate). The question arises as to what data could or should be fed to the model in order to make valid predictions about the state of the driver and his or her ability to take over the driving task, and how the model can be validated if the variable of interest, the implicit understanding of the situation, is not measurable.

## 3.11 Hybrid models – integrating Machine learning approaches into cognitive models

*Moritz Held (Universität Oldenburg, DE)*



A key step to understanding driving behavior are the underlying cognitive processes while driving. With the help of computational cognitive models, we aim to not only better understand points of human failure but also provide theory-driven predictions for different driving scenarios. Combining theory-driven white box models like ACT-R models with data-driven machine learning techniques ("black box models") is an intriguing approach towards creating predictive models which are usable in autonomous vehicles. In the Dagstuhl Seminar we identified key areas in which black box approaches are more useful either from a computational point of view (e.g., lower computation time in highly time-critical scenarios) or when it comes to estimating Driver states/traits. On the other end, white box approaches can be useful when the model needs to be robust to handle unseen scenarios or when the aim of the model is to understand the cognitive processes responsible for the behavior. Hybrid models, which integrate machine learning techniques to restrict the behavior of a cognitive model seem promising.

## 3.12 The Nomadic Worker: Autonomous Vehicles as Future Worksites

*Shamsi Tamara Iqbal (Microsoft – Redmond, US)*



With a rapid transition to a new future of work, the traditional definitions of where, how and when people can get their best work done is continuously evolving. An important finding of the years of the pandemic showed that people appreciated not having to commute as they had to work remotely and found that they could rather use the time previously spent on commute more productively while working from home. Yet with hybrid work practices, many workers will be returning to some form of commute at least during some of their work week. A key research question is how workers can effectively use the commute time – be it for their productivity or personal wellbeing needs without jeopardizing driving safety. In hybrid work scenarios, this becomes an even more important question, as efficient use of time for personal and collaborative productivity and wellbeing will be paramount for workplace performance. I propose a research agenda looking at the challenges of making the car a temporary worksite, where people can safely and comfortably get work done or attend to their personal needs without the worry of "wasting time in commute". I envision three stages – 1) fully manual (present day), 2) Level 3 where the driver is still in charge of driving with some autonomy support from the car, 3) fully autonomous – and the nature of work that people can get done will be very different from one stage to the next. The value of having three stages is that we can start experimenting right away and inform vehicles of the future of the general challenges of working in the car – limited attention span, environmental constraints such as motion sickness, lack of large workspaces – and extend and adapt to the unique needs of a particular stage. This work cross cuts both the domain of Future of Work, as well as Autonomous Vehicles and will facilitate innovations of how to get work done in non-traditional worksites.

## 3.13 An exciting time for the field of Cognitive Modeling of Human-Automated Vehicle Interaction

*Christian P. Janssen (Utrecht University, NL)*


**Main reference** Christian P. Janssen, Linda Ng Boyle, Wendy Ju, Andreas Riener, Ignacio Alvarez: "Agents, environments, scenarios: A framework for examining models and simulations of human-vehicle interaction", Transportation Research Interdisciplinary Perspectives, Vol. 8, p. 100214, 2020.
**URL** https://doi.org/10.1016/j.trip.2020.100214

It is an exciting time for the field of computational cognitive models of automated driving, and that excitement is palpable at this Dagstuhl meeting — even when attending it remotely. The field of cognitive modeling has been around for at least five decades (depending on how you count, but for example going by this commentary of Allen Newell [1]), and has seen a lot of progress. The last few years have seen a growth in the types of AI and modeling techniques that have been developed for and in Human-Computer Interaction [2]. Moreover, there is excitement to include these techniques in the Automotive domain [3]. A challenge is though that with the breadth of techniques, and the breadth of application areas it is easy to get lost in translation. Following a preceding Dagstuhl [4], we had already identified that it is valuable to distinguish between simulations of the agent, its environment, and scenarios [5]. Similarly, at this Dagstuhl we had various discussions on how to push the field forward: refine and integrate techniques for modeling into the automotive domain. My hope and expectation is that it will be a win-win-win situation. First, through modeling the automotive field will gain better insights into its users and can better design for it. Second, the social sciences and related fields will have an area to test their models, to see if they also work in more practical sessions and to identify where refinement is needed. Third, through this endeavor a solid scientific community will form. I can't wait to see what the future brings!

### References

1 Newell, A. (1973) You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In Chase, W.G. (Ed.) *Visual Information Processing*. New York: Academic Press.
2 Oulasvirta, A. (2019). It's time to rediscover HCI models. *Interactions*, 26(4), 52-56.
3 Jeon, M., Zhang, Y., Jeong, H., P. Janssen, C., & Bao, S. (2021). Computational Modeling of Driving Behaviors: Challenges and Approaches. In *Extended Abstracts of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 160-163).
4 Riener, A., Boll, S., & Kun, A. L. (2016). Automotive user interfaces in the age of automation (Dagstuhl Seminar 16262). In *Dagstuhl reports* (Vol. 6, No. 6). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
5 Janssen, C. P., Boyle, L. N., Ju, W., Riener, A., & Alvarez, I. (2020). Agents, environments, scenarios: A framework for examining models and simulations of human-vehicle interaction. *Transportation research interdisciplinary perspectives*, 8, 100214.

## 3.14 Technical capabilities for computational models

*Myounghoon Jeon (Virginia Polytechnic Institute – Blacksburg, US)*



Computational models are used to simulate and predict human behaviors by quantifying the relationship between research parameters and behavioral outcomes. In this process, precision of prediction and efficiency of the modeling process might be necessity for modeling, rather than capabilities. Once these fundamental components are satisfied, we can consider adaptability or expandability of modeling, which we can call compatibility. The first technical capability the model should have for compatibility would be sensitivity. It is about whether the model can be sensitive to changes in resource demand. The next one is selectivity. It is about whether the model can be sensitive only to differences in specific resource demand. For example, if task difficulty can be resolved by multiple resources (e.g., secondary task while driving), but the model has only visual components, it might not be able to describe the entire human behavior. Then, the research question would be whether the model can be expanded by having other resources, such as auditory (speech and non-speech) components. The next capability for compatibility would be diagnosticity. It is about whether the model can indicate when human behaviors vary and can indicate the cause of variation. For example, the cognitive architecture framework may not capture other types of human states–e.g., emotions, fatigue, trust, mind wandering while driving. So, the research question would be whether the model can be combined with other constructs to predict and explain other than cognitive constructs. Finally, in terms of precision of the model, as long as the model can postulate the same stimulus-response mapping, it should work. This is the original meaning of ecological validity. Therefore, the simulation does not necessarily have high fidelity software and hardware compatibility (which is more about external validity), but the important question would be whether the model can have the psychophysical similarity. These technical capabilities (sensitivity, selectivity, diagnosticity, and ecological validity) will be useful to assess the compatibility of computational models in future.

## 3.15 HMI design for autonomous vehicles: methodologies and intercultural analyses

*Xiaobei Jiang (Beijing Institute of Technology, CN)*



To ensure driving safety, efficiency, and comfort in mixed traffic environment with autonomous vehicles (AVs), these vehicles must be able to understand the plans and states of partner road users, and can communicate and coordinate their actions with them. With the increasing number of AVs being tested and operating on roads, external HMI (eHMIs) are proposed to facilitate interactions between AVs and other road users. Taking eHMI for crossing pedestrians as an example, many methods like user interview with images & videos, Wizard of Oz (WoZ), virtual reality, the Delphi method, on-road experiment, etc have been conducted in studies focusing on the effect of eHMIs. But different methods are subject to different biases, even yielding conflicting results. This problem should be carefully considered. As there are cultural differences in the way how goals and states are expressed by HMI as well as how HMI is interpreted by others, it is important to understand these differences to allow the integration of AVs into different cultural contexts.

## 3.16 Benefits and Challenges of Computational Cognitive Models

*Jussi Jokinen (University of Jyväskylä, FI)*



The benefit of using computational cognitive models is that they force the modeller to make theoretical and practical assumptions explicit. For instance, in the case of semi-autonomous vehicles, understanding how the human driver adapts to the presence of different automatic driving assists helps to design safer and efficient driving. Moreover, models can be used to generate predictions of how drivers adapt to various design choices. However, modelling is time-consuming and hard, and involves a learning curve that may be unacceptable. In order to facilitate proliferation of computational cognitive modelling, the modelling workflow must be made usable. This involves creating tools for modelling, but also benchmarks for model testing, and "canonical scenarios" for trying out design and modelling ideas in familiar environments.

## 3.17 Much more data is needed

*Wendy Ju (Cornell Tech – New York, US)*



To build computational models for human-AV interaction, we need a lot more data. Of course, we have all heard that "all models are wrong, some models are useful." To date the models I have seen in the space of human-AV interaction do not account for the cultural variation, contextual variation and temporal variation that we know for certain would affect any predictions for how people will respond to an AV.

### Instrumentation

One issue with gathering data to address cultural and contextual variation, at least, is that is is unlikely that any one research group in one location would be able to gather data across two sites to even begin to address this issue. For this reason, the question of publishing and sharing the way that instrumentation occurs in lab and real-world data collection. Sharing instrumentation set-ups makes it easier to have datasets that are comparable; for this reason, publications that discuss and benchmark different instrumentation configurations for data collection of human-vehicle interaction should be considered in and of themselves to be contributions to the community.

### Scale

The scale of the data that would need to be collected to address cultural, contextual and temporal variation is also an issue. Finding ways to use machine learning to augment and scale human observation and coding in empirical data analysis is critical. As a community, we should be discussing computational methods to share scale and validate data analysis.

### Data

I believe that part of what is needed is that we as a community need to treat the gathering and sharing of data as a real contribution, even before it has been analyzed or yielded any insights. Empirical data, particularly from real-world studies, is often "dirty"– equipment failures, participant strangeness, and exceptional events abound. Given the sparsity of of this much needed data at all, it is more important that any issues with the data be clearly documented and explained than that we demand great perfect data that fits our models.

## 3.18 Computational cognitive models for attention monitoring

*Tuomo Kujala (University of Jyväskylä, FI)*



Because of the complex and unpredictable nature of traffic with human road users, I argue that in order to be safe, driver should monitor the behavior of the car and the traffic environment up to SAE level 3 of automated driving (maybe even at level 4). This requirement is decreased as the automation level increases and that is actually a big problem as the driver's attentional capacity is freed but not fully. SAE levels are also somewhat problematic in this sense as a real car given a level 3 status may not always function as a level 3 car should (e.g., always know when the driver should be alerted to take over).

Current driver attention monitoring systems are really impressive in their object and behavior classification performance but I would argue that it is not enough to monitor the driver or insides of the cabin to know if the driver is inattentive towards driving. Drowsiness is probably the only form of inattention that can be reliably detected by monitoring the driver only. The attentional demands of driving vary based on traffic situation, surroundings, upcoming situation, etc. and there is often so-called spare attentional capacity in driving. And even more when the level of automation increases. But how much, that is the question? Definition of these requirements might also get harder when we go up in the level of automation. As this task is fairly complex, we would need computational models for this.

For these reasons, and because driving is a safety-critical context, we'll need also prescriptive computational models, not only descriptive and predictive. For instance, we would need a normative criterion to define when the driver is inattentive, depending on situational and driver-specific variables

## 3.19 In-vehicle human-automation interaction: Bumper-to-bumper traffic, and short bursts of activity in non-driving tasks

*Andrew Kun (University of New Hampshire – Durham, US)*


**Main reference** Divyabharathi Nagaraju, Alberta Ansah, Nabil Al Nahin Ch, Caitlin Mills, Christian P. Janssen, Orit Shaer, Andrew L. Kun: "How Will Drivers Take Back Control in Automated Vehicles? A Driving Simulator Test of an Interleaving Framework", in Proc. of the AutomotiveUI '21: 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Virtual Event / Leeds, United Kingdom, September 9-14, 2021, pp. 20–27, ACM, 2021.
**URL** https://doi.org/10.1145/3409118.3475128

A driving scenario that promises to be important for human-automation interaction in vehicles in the near term is bumper-to-bumper traffic. In this scenario the vehicle is relatively easy to control by automation, and there is a relatively low risk of injury in case the automation makes a mistake. At the same time there would be a large positive impact on the driver if they could use the time that the vehicle is in bumper-to-bumper traffic to engage in non-driving tasks.

How can drivers use their time if the car can drive itself in bumper-to-bumper traffic? The answer is: short bursts of non-driving activity with rapid returns to driving. The reason is that bumper-to-bumper traffic doesn't last forever – it's likely to often take only minutes. And transitions back to driving will need to be very quick, on the order of seconds, not minutes [1].

It is likely that those short bursts of activity will very often involve manual-visual tasks. We recently conducted a time-use study with 400 knowledge workers who commute by driving [2]. We asked them what they would like to do in a future, safe automated vehicle. They provided us with tasks they are interested in, and we assessed the tasks in terms of the need for various cognitive resources (using the Wickens multiple resources model). We found that our participants wanted to do more manual-visual tasks, both for work and for personal tasks. Thus, we expect more typing and browsing, in contrast to silent reflection or listening to music.

### References

1 Nagaraju, Divyabharathi, Alberta Ansah, Nabil Al Nahin Ch, Caitlin Mills, Christian P. Janssen, Orit Shaer, and Andrew L. Kun. (2021) How Will Drivers Take Back Control in Automated Vehicles? A Driving Simulator Test of an Interleaving Framework. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 20-27).

2 Teodorovicz, Thomaz, Andrew L. Kun, Raffaella Sadun, and Orit Shaer. (2022) Multitasking while Driving: A Time Use Study of Commuting Knowledge Workers to Assess Current and Future Uses. *International Journal of Human-Computer Studies.*

## 3.20 Four challenges to computational models in Human-AV interaction

*Dietrich Manstetten (Robert Bosch GmbH – Stuttgart, DE)*



Computational models (including but not limited to cognitive models) are playing an essential role to enable safe automated driving and to make it a reality on tomorrow's roads. We are looking at four challenges in the field.

1) **Modeling driver engagement and driver availabilty.** Driver monitoring becomes a must in the homologation of assisted and automated driving. We need real-time models estimating whether the driver fulfills his supervision task in Level 2 (engagement) and whether he will be able to take over if a request occurs in Level 3 (availability).
2) **Take-over performance models.** Respecting the current status of traffic, vehicle, and driver we need models to predict the temporary and quality aspects of the take-over activity and to decide if his/her actions are intentional and supportive for safe driving.
3) **Shared control / cooperative driving.** As humans are still needed for subtasks in automated driving (e.g. for deciding on lane changes and other tactical aspects)driver's computational models should be used during system development to analyze and validate the successful cooperation between driver's actions and automation control.
4) **Good driving behavior as a role model for automation.** Behavioral models describing the human driving process in traffic are a good way to serve as a role model and to be mimicked by future automated and even autonomous driving systems. Separating "good" and "bad" driving can help that autonomous driving will be able to realize a chauffeur-like safe and comfortable driving style.

## 3.21 Computational models of humans as tools for enabling safe and acceptable vehicle automation

*Gustav Markkula (University of Leeds, GB)*



Computational models of human cognition and behaviour have applied uses in, and in some cases will be indispensable for, development of safe and human-acceptable vehicle automation. At lower levels of automation (e.g., SAE level 2-3), models of how humans monitor the automation and take, receive, or share the vehicle control can be used (1) as part of online, real-time algorithms to infer driver state and predict driver actions, and to adapt vehicle behaviour accordingly, or (2) in offline computer simulations to evaluate different automation design alternatives. The same division into use in both online algorithms and offline testing applies also for higher levels of automation (e.g., SAE level 4-5), but then with respect to modelling the cognition of behaviour of other road users around the automated vehicle. One exciting prospect and grand challenge, across all of those application areas, is to find the balance and combinations between mechanistic/cognitive models on the one hand, and data-driven/machine-learned models on the other, to enable extensible models that users (automated vehicle designers/engineers) can generalise to new contexts and data sets, without direct involvement from the original model developers.

## 3.22 Data for and models on Trained Operators

*Nikolas Martelaro (Carnegie Mellon University – Pittsburgh, US)*



Current thinking around human-machine interaction in autonomous vehicles is often through the lens of untrained everyday users. We may often think that our models of human behavior are generalizable. However, after training, people may develop new ways of acting, especially with autonomous systems. ,A potential area of work that is underexplored in the space of autonomous vehicles is how trained operators will act. For example, a bus operator or a semi-truck driver will have more training in both driving and with the systems they are using. As such, cognitive models for these people may be different than those of a lay population. The research community should consider research on trained operators and should capture data in experimental settings and develop models that are specific for them. Such work may lead to better predictive models that can be used in real-world autonomous vehicle sooner than in vehicles with a lay population. This line of research may also identify good mental models that could then be taught to lay drivers to help improve their interactions with autonomous vehicles.

## 3.23 Simulation Intelligence and Computational Models in Autonomous Vehicles

*Roderick Murray-Smith (University of Glasgow, GB)*



"Simulation Intelligence" brings together first principles models, probabilistic programming and machine learning. It has a lot of potential for supporting a more principled approach to creation of computational models, which can help support and direct scientific progress in the field. Moving from correlational fits to data towards causal models which really represent the actual behaviour will be vital for reliable behaviour in novel contexts. Using forward and inverse inference mechanisms, as is common in other areas of science, gives a potentially more reliable way of formulating the problems, and managing models for components of larger models. Closed-loop aspects are also critical – they affect how we acquire data, and also provide challenges associated with the interdependence of human behaviour on the dynamics of the vehicle (e.g. crossover models), but this seems to be underconsidered at the moment.

## 3.24 Computational Rationality as an Emerging Approach to Inform the Design of Interactive System

*Antti Oulasvirta (Aalto University, FI)*


**Main reference** Stuart K. Card, Thomas P. Moran, Allen Newell: "The psychology of human-computer interaction", Erlbaum, 1983.
**URL** https://www.worldcat.org/oclc/59953542

How do people interact with computers? This fundamental question was asked by [1] with a proposition to frame it as a question about human cognition, in other words as a question of how information is processed in the mind. Recently, the question has been reframed as a question of adaptation: how do people adapt interaction to the limits imposed by cognition, device design and its environment? The core assumption of computational rationality is that users act according to what is best for them given the limits imposed by their cognitive architecture and their experience of the task environment. The theory can be expressed in computational models which explain and predict interaction and be therefore used for design and adaptive systems.

**References**

1 Card, S. K., Moran, T., & Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: L. Erlbaum Associates Inc.

## 3.25 A common understanding of models...

*Andreas Riener (TH Ingolstadt, DE)*


**Main reference** William Hudson: "Toward Unified Models in User-Centered and Object-Oriented Design", p. 313–362, Addison-Wesley Longman Publishing Co., Inc., 2001.
**URL** https://www.syntagm.co.uk/design/articles/chapter_09_hudson.pdf

In many discussions I have had over the last years with my students (in the interdisciplinary program User Experience Design), within my research group but also with colleagues (e.g. in the context of Dagstuhl Seminar 16262), it has turned out again and again that there is a different understanding of what a model is and how or for what it can be used. The main reason was that we have different backgrounds (engineering, computer science, design, social sciences, psychology). We need to find ways and means to make sure we ("the system designer") mean the same thing... Predictive models could be a common basis – but of course they are only a simplification of real behavior [1]. If we focus on human-machine (or human-AV interaction), it is particularly necessary that both parties (human, machine) understand each other and can mutually assess how far the partner is trustworthy in a certain situation [3]... And this is particularly important in the recent research field "Cooperative Driving" – where automation and human work together as team players [2]. In a team, each agent must be informed about the other agents' current activities, their strategies, the status of their efforts, if they are having problems, and their intentions for planned actions. The Object, View, and Interaction Design (OVID) model [1] (Roberts, 1998) contains three design models based on Don Norman's notion of cognitive engineering [4] and clearly visualizes the problem: We cannot really design a user's conceptual model, as this is highly inter-personal (but also intra-personal), based on experience, and exist only in the brain of a person... Furthermore, user interface and vehicle designers as well as user experience practitioners are often challenged with the question for which user group they are designing for – each with different needs, different interests, and very different ways of interacting with technology. In vehicle production, we do not have the luxury of focusing on only one group (at least not so far), i. e., designing vehicles for specific age groups or cultures, that's why interaction designers and engineers must learn to recognize and reconcile the needs of their main user demographics. This problem will remain even with fully automated driving when using the car as a place for relaxation, entertainment or work. Defining characteristics, differences, and tensions between individual user groups might help to account for different individuals. An important question in this regard is, whether or not there is a single system suitable for all (or at least most) customers and stakeholders. Is it axiomatic to target user groups differently? (HMI and system configuration allow to...). Dagstuhl Seminar 22102 was the perfect place to discuss with all the participants from different backgrounds (my access is hypotheses-driven experimental research) the possibilities of combining existing models, extending/improving them, or developing new models to better represent the wide variety of variants in human-AV interaction/cooperation and human diversity. I hope that results can be derived from this week's collaboration that will help the community....

**References**

1 William Hudson (2001).Toward Unified Models in User-Centered and Object-Oriented Design. In *Object Modeling and User Interface Design Designing Interactive Systems*. Addison-Wesley, Pages: 313-362. ISBN: 0201657899 (Figure 9.4; p. 328).
2 Frank Flemisch, Matthias Heesen, Johann Kelsch, Julian Schindler, Carsten Preusche, and Joerg Dittrich (2010). Shared and cooperative movement control of intelligent technical systems: Sketch of the design space of haptic-multimodal coupling between operator, co-automation, base system and environment, Vol. 1.
3 Philipp Wintersberger, Anna-Katharina Frison, Andreas Riener, and Linda Ng Boyle (2016). Towards a personalized trust model for highly automated driving. *Mensch und Computer*, Workshopband, 2016.
4 Norman, D. A. (1986). Cognitive engineering. In D. A. Norman and S. W. Draper (Eds.) *User Centered System Design* (pp. 31–61). Hillsdale NJ: Erlbaum.
5 Martin R. Baumann and Josef F. Krems (2009). A Comprehension Based Cognitive Model of Situation Awareness. In V. D. Duffy (Ed.), *Digital Human Modeling*, Vol. 5620, pp. 192–201, Springer, https://doi.org/10.1007/978-3-642-02809-0_21.

## 3.26 Anticipating the cognitive state of the driver

*Nele Rußwinkel (TU Berlin, DE)*



The next big challenge for computational models of human-Automated Vehicle Interaction is, from my perspective, to integrate more aspects and understanding of the human perspective. I do not think that one modelling method needs to capture all possible aspects of a scenario and related entities to be a good model. What is needed are good models that can capture specific aspects and a good concept that can integrate this information with emerging transparency – how different aspects have led to a specific understanding of the situations. A modular approach might be helpful to achieve this in case some module results have a symbolic form. What is needed is a tighter connection of human and technical system understanding. It is not sufficient to measure some mental states, it is also (in some situations) necessary to understand what the cause of the measured state is and how to address it. I claim for more research on models that anticipate the human driver in the dynamic situation and include mental representations e.g., for "Situation Awareness" or "Mental models" or specific "Expectations" about what the human considered relevant now and what is expected to happen next. This would enable a better mutual understanding and a more natural (less effortful) shared control of the car or the situation and to avoid misunderstanding.

## 3.27 Models that Inform Design

*Shadan Sadeghian Borojeni (Universität Siegen, DE)*



The advances in automation and machine learning are changing the designs of in-vehicle user interfaces from static command-response "buttons & displays" to dynamic adaptive interfaces that adjust themselves to user behavior and preferences. From a modeling perspective, however, this can be a challenge to integrate these new "emerging behaviors" into the models. On one side, having simplistic models allow us to scale them to different scenarios and interaction forms, on the other side, it restricts our ability to precisely describe/predict the adaptive and dynamic interactions between the user and the automated vehicle.

Another challenge is answering the question of whom we are designing for and in which scenario? The first step towards that is defining the sides of the interaction, i.e., the human and the automated vehicle. Identifying the capabilities and responsibilities of each of these partners across different levels of automation can help us understand what factors are required to be considered in the models. For example, for an urgent takeover situation in level 3, factors that define users' situational awareness, the urgency of the situation, and contextual factors might be determinants of the behavior. However, for a level 5 scenario, where the interaction goal is mostly to assure users' comfort and wellbeing, the determining factors stem from users' psychological needs and affective state rather than her situational awareness. It is, therefore, necessary to first define users' tasks at each level of automation and in different scenarios. And then elicit the factors that are required to fulfill these tasks' goals. This allows us to create a modular model of tasks, levels of automation, and required factors that assist us to identify the important elements that should be addressed in the design of in-vehicle user interfaces.

## 3.28 Complexity of computational models

*Hatice Sahin (Universität Oldenburg, DE)*



Computational models can help to ease the cost of experimental studies. Discussions have been developed under the consensus that various individual or external factors affect distribution of driving responsibility between the human and the automated vehicle. These factors range from personal preferences, trust, experience, social and legal norms to environmental conditions. While it is crucial that these factors are recognized, it may not be vital for computational models to include all possible parameters. It is still a matter of discussion whether their benefit is limited by their simplistic qualities. While some might consider that they might have to be complex enough to provide generalizability, they could still be helpful to investigate individual factors with limited complexity.

## 3.29 Teamwork and Driver-Vehicle Cooperation

*Boris van Waterschoot (Rijkswaterstaat – Utrecht, NL)*


**Main reference** Sebastianus Martinus Petermeijer, Angelica M. Tinga, Reinier Jansen, Antoine de Reus, Boris van Waterschoot: "What Makes a Good Team? – Towards the Assessment of Driver-Vehicle Cooperation", in Proc. of the AutomotiveUI '21: 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Virtual Event / Leeds, United Kingdom, September 9-14, 2021, pp. 99–108, ACM, 2021.
**URL** https://doi.org/10.1145/3409118.3475153

With the introduction of driving automation systems [1], the subtasks of the DDT are being performed by the (human) driver, by the driving automation system, or by both. This means that under certain circumstances, human and automation have become collaborative partners in executing (parts of) the driving task.

Although the view of driver and automation acting as collaborative partners has been well established, means to assess their teamwork are still lacking. Moreover, available evaluations are usually addressed either from a technical stance or from a human factors viewpoint, which does not comply with a general acknowledged view of a unified driver-vehicle system. This stance on teamwork and a description of our aim to evaluate driver and vehicle cooperation by means of a framework can be found in [2] .

Our work is currently being followed up by a roadmap – which is shared as soon as possible – aiming to deliver guidelines on the framework's practical implementation, potentially supporting monitoring, evaluation and design activities concerning driver-vehicle cooperation.

This abstract is a call to get involved in our activities regarding the objective evaluation of driver-automation cooperation.

**References**

1 SAE, Society of Automotive Engineers. (2018) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. J3016 JUN2018. SAE International: Warrendale, PA, USA.

2 Petermeijer, S. M., Tinga, A., Jansen, R., de Reus, A., & van Waterschoot, B. (2021). What Makes a Good Team? Towards the Assessment of Driver-Vehicle Cooperation. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 99-108).

## 3.30 From Trust Studies to Trust Models

*Philipp Wintersberger (TU Wien, AT)*



Six years have passed since my first Dagstuhl Seminar. At that time, the first deadly crash with an automated vehicle happened. It became clear that the trust relationship between drivers and automated vehicles will become a key issue for guaranteeing safety in the future since drivers are not necessarily automation experts. Thus, we intensively discussed the topic of trust in automated vehicles and proposed to research "personalized trust models" so that in-vehicle HMIs can adapt to drivers' states, strengths, and weaknesses. A great variety of trust studies have been conducted and published since that seminar, revealing many relevant factors influencing this multidimensional psychological construct. However, a sophisticated trust model has still not been developed. Thus, I am very happy for the invitation to this modeling seminar, as I got the chance to (1) talk to modeling experts from my domain and (2) discuss with colleagues concrete, actionable steps to move towards the goal outlined above. I am confident that it will not take another six years to integrate all the past study results and develop an initial functional version of the proposed model.

## 3.31 Understanding and challenges of computional models of human-automated vehicles interaction

*Fei Yan (Universität Ulm, DE)*



In order to better understand the interaction process between humans and automated vehicles as well as develop adaptive assistance systems to support drivers, computational models that can precisely predict driver behaviors are needed. However, with the change of automation levels, there are challenges while developing computational models. With the change of driver's role from SAE L2 to SAE L3, the focus of modeling changes from monitoring driver's states, to modeling driver's takeover reactions which are more safety critical compared to SAE L2. From SAE 4 or SAE L5, the transition of responsibility can be relevant. While modeling takeover process from SAE L3,indicators like takeover timing as well as takeover quality such as maximal acceleration and TTC, subjective measurement such as workload can be taken into account. Facing the complex scenario in the mixed traffic, the combination of cognitive approach with black-box models is needed, which can maximize the advantages of different models, but also ecologically save the cost of developing complex cognitive models. The cognitive theory, evidence or empirical research should be made use of to help to develop casual black-box models, which can help to interpret the relations between relevant factors.

# 4 Working groups

## 4.1 Learning and Adaptation

*Jelmer Borst (University of Groningen, NL), Alexandra Bremers (Cornell Tech – New York, US), Birsen Donmez (University of Toronto, CA), Mark Eilers (Humatects – Oldenburg, DE), and Roderick Murray-Smith (University of Glasgow, GB)*



Learning was discussed from two points of view: on the one hand users have to adapt to new (autonomous) systems, and on the other hand models need to be able to learn and adapt.

With respect to users, they have to learn operating (semi-)autonomous vehicles, and also adapt to such vehicles being present in the environment. One challenge is that different systems (e.g., Tesla, GM) might behave differently, which makes learning harder. In addition, it is often not clear whether they operate in autonomous mode or not. Second, people might "abuse" the new systems, for example crossing the road directly in front of an autonomous car, as it will stop anyway.

Models will also have to adapt to new situations on the road (i.e. more autonomous cars) and to adjusted human behavior. We can differentiate between quantitative changes (new speed limit) and qualitative changes (new traffic rules). While the former is probably possible, the latter is much more difficult to automatically adapt to. In general, we expect that adaptation, in particular qualitative, is easier for cognitive and causal models, as these take content into account.

To adapt machine-learning-based models, one could think of using feedback of the user. However, some drivers might be more amenable to give feedback than others. We therefore recommend implicit learning, based on the behavior or physiological responses of the users.

## 4.2 How can model use be increased in design?

*Alexandra Bremers (Cornell Tech – New York, US), Jussi Jokinen (University of Jyväskylä, FI), and Gustav Markkula (University of Leeds, GB)*


**Joint work of** Bremers, Alexandra; Jokinen, Jussi; Markkula, Gustav

There are examples of models used by designers and engineers in the industry, such as for traffic testing. However, a question remains about how accessibility can be improved, and adaptation of models can be increased.

Potential roadblocks and limitations for using models in design are:

1. Cost-benefit tradeoff. An investigation is needed to clearly map out scenarios where modelling adds the most value in addition to testing methods.
2. Design space restriction due to specificity of models. For instance, a typing model would assume touch-based interaction within a specific limited interface size.
3. Language intersection between the designer's creativity and the model's formalized way of using specifications. In machine learning, tools have been developed, such as the OpenAI gym, which bridges the gap between model and application.
4. Complexity depending on the task. Tasks that seem simple from a design perspective, such as selecting the best option in a webshop, could involve a very complex cognitive model.

In AV applications, HMI and UI design seem ripe for model integration. Successful tools exist, such as Distract-R and OpenAI Gym, online design tools such as the Adobe Mixamo library, and wireframing tools like Miro and Sigma. Combining these could result in a tool for both quick model-assisted sketching and a more thorough model-based evaluation. The behaviour of other road users, such as children playing soccer on the sidewalk, could be a lot more complex to realize. Discussions with AV industry engineers could further confirm specific needs and requirements.

## 4.3 How can empirical research support computational models (and vice-versa)?

*Debargha Dey (TU Eindhoven, NL), Moritz Held (Universität Oldenburg, DE), and Fei Yan (Universität Ulm, DE)*



In this focus group, we dove deep into the caveat of empirical research and computational modeling approaches and discussed the situations where each is applicable, and where one approach can benefit from the other. While a modeling approach is not needed to answer every research question, it can be critical if we are trying to justify cognitive processes (white box models), or trying to reduce the complexity of the research space by looking at causality and/ or relationship between certain parameters of the environment and the observed behavior (black box models). We recognized that there can be no one strict guideline to the question of the data points needed to develop a model (as it largely depends on the research question). In situations when it is difficult to estimate the impact or relevance of certain parameters for heuristic approaches, modeling approaches can help by highlighting them. Thinking through the cognitive architecture of a white-box model can identify the "independent variables of interest" for empirical studies. Furthermore, data from empirical research can be helpful in building hypothesis for computational models especially predictive models using machine learning techniques. However, "absolute" black box models lack the transparency and explainability in terms of causality explained through cognitive processes, which calls for inputs from empirical research and cognitive theory towards a combined approach that leverage the best of each world.

## 4.4 Modeling Long-term Effects in Human Technology Engagement

*Patrick Ebel (Universität Köln, DE), Hatice Sahin (Universität Oldenburg, DE), and Philipp Wintersberger (TU Wien, AT)*



We discussed how long-term effects (distribution shifts) in human behavior when interacting with technology can be modeled.

To make the problem more tangible, we have decided on a specific use case: "How to model the probability for a driver to activate automated driving functions?" Such an approach could be an interesting element for driver monitoring systems, since usage patterns can be related to interaction problems (such as trust) [1]. Given a certain set of input variables (environmental context, individual driver characteristics, trust in the system, effect of "events", etc.) the model should predict how likely it is for the driver to activate driving automation. The goal of such an approach is to (1) model long-term driver behavior and to (2) develop effective measures to increase the usage of automated driving functions.

To model the probability of human engagement with technology, three main components are be modeled over time:

1. The general probability of engagement. This probability is dependent on factors like the benefit of using the technology (How good is the driving automation?), learnability (How well do drivers adapt?), trust (Do Drivers trust automation?), individual characteristics, and the probability of being in a situation where the technology is applicable (is it possible to activate automation in a specific driving situation?). This probability can be modeled as a continuous function over time and serves as the base probability of engagement.
2. The influence of incidents. We defined incidents as event-based disturbances that influence the overall probability of engagement at a certain point in time. They can be modeled using a step function. An example would be the sudden change of the activation probability after experiencing a critical situation while driving with automation activated. A relevant question in this regard is, if the underlying cause for a behavior change can be derived from a changing usage pattern (i.e., do there exist detectable patterns for certain events)?
3. The effect of interventions. An HMI intervention is provided to a human to mitigate the effect of an incident. Interventions can be modeled by adding an "intervention effect" function to the general engagement function. The idea is that a positive intervention like explaining why a critical situation couldn't be solved by the automated driving function can change the probability of engagement over time such that it approaches the original probability before an incident happened more quickly.

Subsequently, we developed a first sketch for a study to investigate the three main components of the theoretical model. This study needs to be further developed.

**References**

1 Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.

## 4.5 Focused session "Application in specific scenarios"

*Martin Fränzle (Universität Oldenburg, DE), Luisa Heinrich (Universität Ulm, DE), and Roderick Murray-Smith (University of Glasgow, GB)*



As applications and application scenarios make their way into the research stream mostly via use cases and benchmark examples, the discussion in the group focused on how to obtain representative domain coverage via benchmarks and use cases. The lack of a reasonably broad and generally accepted set of such benchmarks and use cases was identified as an impediment both to rigorous and comparative assessment of the current state of research and to the identification and prioritization of open research problems. An informed, problem-driven advancement of the state of the art would thus require to find, collect, and publicize benchmarks that collectively obtain representative coverage of the relevant applications and application scenarios and that facilitate rigorous evaluation of and friendly competition between modelling approaches.

This quest in turn induces the need for an intelligible characterization of benchmarks that permits a mapping of individual benchmarks as well as benchmark sets concerning their particular contributions, thereby supporting demand-driven selection and adoption of existing benchmarks. For this mapping of the area, we came up with the following dimensions and explications of dimensions:

1. The dimension of paradigm of use of the human models, distinguishing between human models as proxies of human behavior in model-based design and the embedding of human models into an operational system.
2. Means of parametrisation and validation, namely observed in real-life / on real road / by physical test track driving or through simulator studies.
3. Focus of the embedded human model, ranging over driver only (possibly including in-car passengers also), other traffic participants (in alter cars, in environment – especially VRUs), or both, including reflection of mutual reactive behavior.
4. A broad set of quality criteria for the benchmark specification, including among others heterogeneity of driver behavior and coverage of driver types as well as societal groups (e.g., elderly), non-discrimination, explicit identification of observable/measurable correlates for latent or hidden phenomena of interest, strong correlation of benchmark-represented/benchmark-measurable features to real-world effects and effect strengths, accurate coverage of variability, not just of nominal behavior, as well as means for automated variant generation to avoid overfitting to particular benchmarks, test-retest reliability, inter-rater reliability, and repeatability and reproducibility.
5. Aim of benchmark, covering a wide range of attributes from incentivizing research through friendly competition to facilitation of rigorous relative evaluation and ensuring coherency of methods, quality criteria, etc. between scientific communities. The last dimension identified was 6. technical prerequisites like executability on common hard- and software platforms, convenient packaging (e.g., via containers), and issues of open access and open science.

The discussion group holds the belief that such a classification would enable consolidated and collaborative research, stabilize and align the pertinent quality criteria across domain-relevant disciplines, and thereby advance scientific progress. Necessary first steps to render it reality would be to, first, generate a questionnaire permitting individuals and research groups to locate their respective benchmarks in the above multidimensional criteria set, second, to establish a platform for publication of benchmarks, including video tutorials and supportive contacts, and, third, to foster industrial contribution.

## 4.6 Towards common tests for Models

*Christian P. Janssen (Utrecht University, NL), Jelmer Borst (University of Groningen, NL), Benjamin Cowan (University College – Dublin, IE), Justin Edwards (ADAPT Centre – Dublin, IE), Mark Eilers (Humatects – Oldenburg, DE), and Jussi Jokinen (University of Jyväskylä, FI)*


**Main reference** John R. Anderson, Christian Lebiere: "The Newell Test for a theory of cognition", Behavioral and Brain Sciences, Vol. 26, pp. 587–601, 2003.
**URL** https://doi.org/10.1017/S0140525X0300013X

We discussed whether there can be a benchmark test, or evaluation method, to better compare different models. We were inspired by how competitions are done in the language community. In that community, there are multiple competitions and test frameworks. Not any of those is the "killer test", but across tests one can see where a model or framework is stronger or relatively weaker. This moves the community forward.

In the field of computational modeling for human-automated vehicle interaction there are perhaps two challenges:

1. What is a good model that can perform well on an evaluation test (or tests)?
2. What is an appropriate (large enough, diverse) dataset to test these models on?

We foresee that this can be approached in three steps, that each require attention:

1. Identify what the possible criteria are to evaluate a model on and define what scenarios are needed to test these.
2. Collect data in an empirical scenario where the above features emerge
3. Develop models that can then be tested on this dataset / scenario.

In later discussions we delved deeper into question 1. As a constraint we set that the model should exhibit human-like behavior. We looked at other papers that have defined model criteria, in particular papers by Anderson and Lebiere [1] and Taatgen and Anderson [2]. For these papers, we went through the criteria they defined and determined if and how they apply to the driving domain. We've identified that some modifications and specifications are needed. We also foresee that additional criteria – such as the ability to perform driving tasks and driving specific tests in a human way – need to be added. We plan to further refine these ideas in a follow-up project.

### References

**1** Anderson, J. R., & Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral and brain Sciences*, 26(5), 587-601.
**2** Taatgen, N., & Anderson, J. R. (2010). The past, present, and future of cognitive architectures. *Topics in Cognitive Science*, 2(4), 693-704.

## 4.7 What are important scenarios?

*Dietrich Manstetten (Robert Bosch GmbH – Stuttgart, DE), Moritz Held (Universität Oldenburg, DE), Nele Rußwinkel (TU Berlin, DE), and Fei Yan (Universität Ulm, DE)*



In the beginning of the discussion, it became quite obvious, that we can't write down all important and relevant scenarios. We searched for a framework-like setup, where a scenario is composed of a multitude of independent factors. Some sort of building blocks to construct a specific scenario.

Several of these factors are influenced by the environment and the actual situation. They are relevant, but not directly linked to the automation. We identified

- Traffic environment. Freeway (number of lanes, exit section, tunnel, construction site) / Rural road (curvature) / Urban traffic (crossing, pedestrian, cyclist, parked vehicles)
- Criticality of situation (mostly described by dynamic behavior of others)
- Vehicle Type (truck, passenger car)
- Driver characteristics (stable factors as age / driving experience / knowledge about automation, dynamic factors as time pressure / fitness level)
- Boundary conditions (weather, daylight/night, road surface)

When it comes to human-AV interaction, the main ingredient of the scenario comes from the automation itself. There is the temporary aspect of the current phase of automation, which can pose specific questions: activating the automation, during automation, transfer-of-control. Within these phases, again we see multiple factors, and we partly need to separate carefully between the levels of automation, as some of the effects are relevant for a specific level only.

- Level 2 asks for continuous driver engagement and anticipatory behavior. The transfer of control is usually initiated by the driver.
- Level 3 needs driver availability during automation. Fast set-up of situational awareness after a take-over request. Assessment by take-over time & quality (including safeguard, e.g. mirror glances).
- Level 4 changes the time scale more to minutes than seconds. Some requirement of perceptibility. On system level, there may even be the need for a driver lock-out avoiding unintentional or undesired actions.
- For level 3 and 4 reaching the ODD limit or possible system failures becomes most relevant situation.
- Information design and communication on all levels (dependent of specific situation)
- Clear definition of visual behavior requirements (level 2), cognitive requirements (level 2 and level 3)
- Mode shift with transfer between levels (e.g., from level 3 to level 2) and corresponding challenges of mode awareness
- Aspects of trust towards the system.

The presentation of the group's work in the plenum added the question, where models can be part of the system. This augments the perspective, that knowledge represented in driver models comes partly as a requirement from legal perspectives and regulations. Consequently, the constraints from a legal perspective have to be respected carefully.

## 4.8 What is needed for different SAE level?

*Dietrich Manstetten (Robert Bosch GmbH – Stuttgart, DE), Martin Baumann (Universität Ulm, DE), and Shadan Sadeghian Borojeni (Universität Siegen, DE)*



In this focused session we discussed, which specific system design possibilities are offered within the different SAE levels of automation. We came out with a sketch of a framework to clearly describe the user tasks in the different levels. This description has the potential to serve as a backbone for the system design.

With respect to the levels, we concentrate on the levels 2, 3 and 4. By this, we are adapting somewhat the terminology used by the German Federal Highway Research Institute BASt of assisted (level 2 and below), automated (level 3) and autonomous (level 4 and above) driving, see `https://www.bast.de/EN/Automotive_Engineering/Subjects/f4-user-communication.html`. The time scale for the user's tasks varies as well within these levels, from below seconds (level 2) via few seconds to minute (level 3) to minutes (level 4).

During the automation phase, the main task is shaped by monitoring. Monitoring means to monitor the environment, to monitor the system, and to monitor yourself. The monitoring serves the goal to decide whether I (as the driver) need to intervene, which would end the monitoring phase, and replace it by an action.

We systematically added the view for the phase transitioning from on to off. The possible interventions can be differentiated as an adjustment of the automation behavior, a system-initiated taking back of the control (as a result of a request-to-intervene), and a driver-initiated taking back of the control (might be necessary for safety reasons or just a current driver's preference). As during the automation phase, depending on the level of automation, the concrete task description varies again. We are explicitly looking at the "what?", "why?" and "when?" of the transition.

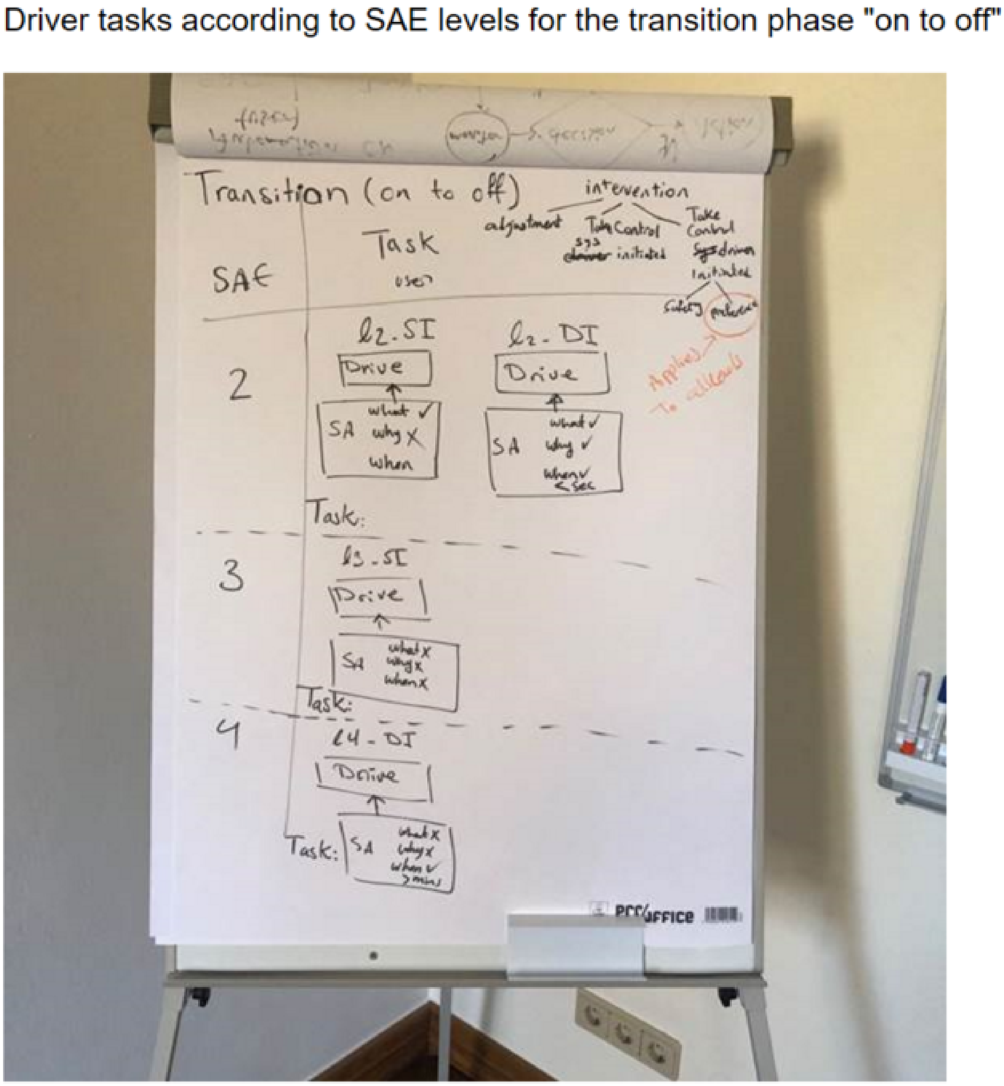
The two main pictures compiled by the working group are added to this report. The group had the clear impression to have a solid ground for an overall framework, but being clearly away from being complete. The focus was so far from the user's perspective only, the system's perspective needs to be added. The group plans to pick it up at a later stage, and to compile a full view out of it. Replacing detailed settings in scenarios by a comprehensive task-oriented view. To be continued.

## 4.9 Multi-agent modelling: Platooning use case

*Gustav Markkula (University of Leeds, GB), Lewis Chuang (LMU München, DE), Debargha Dey (TU Eindhoven, NL), Patrick Ebel (Universität Köln, DE), and Christian P. Janssen (Utrecht University, NL)*

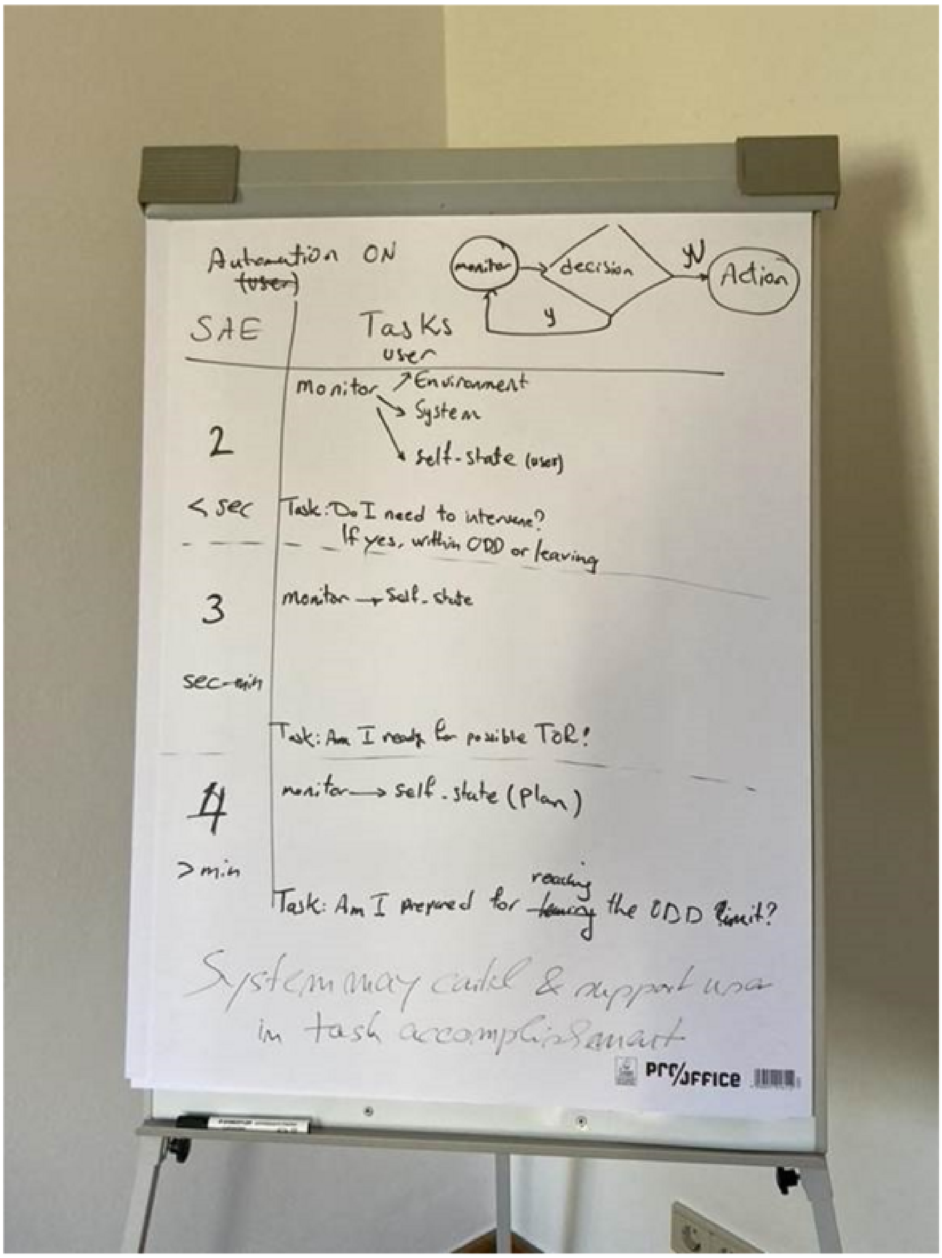


In this focused session, we discussed a number of different scenarios involving multiple human road users that could benefit from the application of human behaviour models. Our discussion converged towards a scenario, centered on platooning – wherein a number of vehicles (cars and/or trucks) follow each other on a highway, at close distance in automated mode, typically




with the expectation that drivers are available to resume vehicular control upon demand. This rich setting gave rise to a number of different control transition scenarios. In particular, we discussed the requirements for enabling the platoon to temporarily break apart to change lanes – e.g., due to an upcoming lane closure – and reassemble later. After discussing the potential contributions of different types of human behaviour models, we agreed on three main types of uses: First, models of the drivers in the lead and following vehicles can inform online algorithms in the platooning vehicles. This would guide real-time decisions how to break up the platoon, by predicting likely human responses to different alternative vehicle actions, and by choosing the control approach that is predicted to give the best outcome in terms of safety, efficiency, and comfort. Second, similar (but possibly not identical, since they would serve slightly different purposes) driver behaviour models could be useful in offline simulations. These would assist an engineer in designing systems by providing possible platoon breakup scenarios for automation implementations and testing – for example, to test lead time requirements for breaking up the platoon. Such tests would optimize safety and comfort, as a function of platoon speed, traffic density, etc. Third, we discussed that

Driver tasks according to SAE levels for the automation phase




models could be useful on a conceptual level. Such models would help system designers and decision-makers better understand the behaviour and limitations of human drivers across different situations.

## 4.10 What computational models are available for human-automation interaction?

*Gustav Markkula (University of Leeds, GB), Debargha Dey (TU Eindhoven, NL), Moritz Held (Universität Oldenburg, DE), Nele Rußwinkel (TU Berlin, DE), and Fei Yan (Universität Ulm, DE)*



In this focused discussion we attempted to provide as broad as possible an overview of currently existing computational models which have been applied to the study of human interaction with automated vehicles, or could be applied in such a context. We listed a large number of references (which will be included in a separate document by the seminar organisers), and as we did so we structured the existing work in a rough taxonomy including the following dimensions: (1) Type of traffic scenario addressed, e.g., automation control transitions, shared control, multi-agent interaction, teleoperation, or driver support systems. (2) Type of (intended) use of the model, e.g., use of models to conceptually guide automation design, use in online algorithms for prediction or decision-making as part of the automation itself, use in offline simulated testing of automation, or as human safety benchmarks for automation. (3) Aspect(s) of human cognition/behaviour modelled, e.g., vehicle control, visual behaviour, individual characteristics, or unobservable mental/cognitive states. (4) Type of model, e.g., machine-learned, cognitive architecture, control-theoretic, and models focused on decision-making. These dimensions are to a large extent independent, but we also discussed that especially different model types, not least distinguishing between machine-learned, black-box models and mechanistic, white box models, might map more or less naturally to different parts of the other dimensions in this taxonomy. We discussed the strengths of these different types of models with respect to time critical decisions, understanding, drivers mental states or cognitive states, subjective and objective complexity, flexibility, driver prototypes and traits. For example, if conceptual understanding is important for the application at hand, white box models have a clear advantage, whereas for example when the complexity of the scenarios that need to be addressed grows, black box models, which can be learned from large naturalistic datasets, may be more appropriate.

## 4.11 Modeling workflows

*Antti Oulasvirta (Aalto University, FI), Alexandra Bremers (Cornell Tech – New York, US), and Tuomo Kujala (University of Jyväskylä, FI)*



In this focused session, we discussed the need to develop modeling workflows specific for this domain. In contrast to empirical sciences (e.g., [1]), models in human-AV interaction need to be, eventually, deployed in interactive systems or as decision-support systems in design. Applicability gains emphasis. However, despite the practical aims of modeling efforts, it is important to retain a level of theoretical and biological plausibility. It was also observed, that because of the safety-criticality of this domain, simulation studies are preferred prior to empirical studies in the wild.

Based on these and other observations, four design principles for a revised modeling workflow were set out: 1) put iteration to the heart of model development; 2) engage in both simulated and real-world studies, 3) aim for robust, verified modeling outputs, and 4) constant contact with theory (in cognitive sciences, biology, neurosciences).

The attached figure shows the revised modeling workflow proposed by the team.

**References**

1 Gelman, Andrew, et al. *Bayesian workflow*. arXiv preprint arXiv:2011.01808, 2020.

## 4.12 What is prediction?

*Antti Oulasvirta (Aalto University, FI), Jelmer Borst (University of Groningen, NL), Patrick Ebel (Universität Köln, DE), Martin Fränzle (Universität Oldenburg, DE), and Nele Rußwinkel (TU Berlin, DE)*



In this focused session, the topic of how to prioritize possible types of predictions was discussed. Prediction is the essential feature of computational models, and it is essential to move beyond merely isolated tests. (Think: Newell's 1973 paper "You can't play 20 questions with nature and win.").

The group agreed that prediction should not be confused with other uses of models, such as description, explanation, and counterfactual reasoning. These are linked to each other but different. The essence of prediction is that it involves some verifiable statement about a future state or unseen situation, given some observational data. What distinguishes driving as a domain is the need to predict what happens in previously unseen future situations, for example when doing verification of an AI's behavior across different situations, some of which can be very rare.

It was also noted that predictions of driver behavior are needed in different ways in different processes, such as user-centered design, safety engineering, and for AI development. Also, different SAE Levels pose different requirements for prediction; like Level 3 poses a requirement for predicting driver's availability. It was noted that an exercise should be carried out to assess these needs from a cognitive modeling perspective.

It was agreed that there is a multiplicity of predictions needed in this domain. This reflects the fact that computational models are used in widely different driving contexts, with different inputs, purposes, and time-ranges. For example, predicting whether a pedestrian is likely to step on a crosswalk occurs in the timespan of a few seconds, while route prediction occurs in the timespan of minutes to hours. There are micro-level and macro-level predictions that are needed, for example arousal states and states related to social interaction with other traffic users. And these interact. There are predictions of latent states that are not directly measurable but measurable via indirect physiological reaction (skin conductance for example). In real applications, we sometimes need to predict behavioral outcomes in driving, but sometimes we need to estimate a latent factor like the level of workload or stress. The group discussed what would be the most valuable thing to predict. There was some consensus that it should be the future state of the driver+vehicle+scene system within some time horizon. This system includes the driver, the vehicle, and the driving environment including the pedestrians.

The need for academic reseach to consider cross-validation practices was discussed. It was perceived as a harder measure of a model's predictive capability. In statistics-based models, parameters are often fit to the same dataset on which fitness metrics are computed. However, the cross-validation practices that are followed in machine learning may not be directly useful for this domain. For example, we sometimes need to predict what happens with a new driver that has not been included in the training dataset (leave-user-out, or LOU, cross-validation). In many machine learning pipelines, the training dataset might contain samples from each user, thus 'leaking' information in a way that is not plausible from a deployment point-of-view in this domain. It was further remarked that safety-related incidents are rare, which complicates prediction, the acquisition of training/reference data for computational models, as well as validation.

Prediction is almost always done under uncertainty, and uncertainty should be factored in the system's action. Therefore a prediction might be more valuable if it also contains a measure of confidence intervals. The need for a probabilistic representation of outcomes is the greater the further in to the future the prediction tries to reach. However, when it comes to the idea of showing this to the driver, or an uncertainty visualization, the problem is non-trivial. Showing uncertainty to a driver can be confusing, as some earlier studies suggest. On this note, "a prediction is an intervention": showing a prediction to a driver affects the driver's future behavior. In multi-agent systems research, there are some formalisms to describe such interactions, but they are rarely done with computational models of humans.

We also discussed the need for updating the way we do models, i.e .the modeling workflow, with methods like sensitivity analysis and what is called parameter recovery.

## 4.13 Which models should be used and what should be their content?

*Shadan Sadeghian Borojeni (Universität Siegen, DE), Martin Baumann (Universität Ulm, DE), Luisa Heinrich (Universität Ulm, DE), Dietrich Manstetten (Robert Bosch GmbH – Stuttgart, DE), Roderick Murray-Smith (University of Glasgow, GB), and Hatice Sahin (Universität Oldenburg, DE)*



In this focused group, we discussed different types of models and their content. Our understanding of a computational model is an executable, precise model that is defined through an unambiguous formulation (mathematical, logical, code, etc.) and is replicable. Computational models can cover different types of models such as cognitive, psychological, social behaviour. To understand what model to use, we first need to define the purpose of the model, e.g., predictive or descriptive. In modelling interaction between humans and automated vehicles, black-box models seem promising. However, the drawback is the lack of knowledge they provide about causal relationships. Therefore, we require to integrate machine learning approaches with causal models, predict the uncertainty of models, and test them in different contexts to understand their validity. This helps us move from black-box data-based models of behaviour towards causal first-principles models, which can generalise better, due to capturing key aspects of the system.

Another topic discussed in this focus group was predicting human behaviour while they adapt to changes in the design of the system. Predicting these emerging behaviours without currently existing data seems challenging. We proposed applying closed-loop models that allow the integration of the adaptive behaviour of humans. Moreover, descriptive models may help us develop causal models that allow predicting emergent behaviour or future behaviour that we currently do not have a lot of data about.

## 4.14 Which scenarios should be taken into account in computational models of human-automated vehicle interaction?

*Hatice Sahin (Universität Oldenburg, DE), Duncan Brumby (University College London, GB), Jussi Jokinen (University of Jyväskylä, FI), and Shadan Sadeghian Borojeni (Universität Siegen, DE)*



We have started by thinking possible scenarios for each level of automation, and soon we realized that we could not possibly list all the scenarios. To have a better structure of the discussion we have focused on categorizing these scenarios from different angles. This brought up whether a taxonomy regarding possible scenarios existed in research. Eventually, we have categorized factors originated by the user and the external world. This way, we presented 2 x 2 table of factors regarding the user and external world in SAE levels 1-4 and in level 5. On SAE levels 1-4 User factors included purpose of commute, driving experience, fatigue and engagement. Trust has been found related to all levels. Level 5 included personal preferences such as sustainability, costs and favorable routes. External or contextual factors were mentioned in all levels. These included social norms, environmental factors such as light and weather conditions and cultural or regional norms and formal rules.

## 4.15 Modeling Trust in Automation

*Philipp Wintersberger (TU Wien, AT), Martin Baumann (Universität Ulm, DE), Justin Edwards (ADAPT Centre – Dublin, IE), Luisa Heinrich (Universität Ulm, DE), and Tuomo Kujala (University of Jyväskylä, FI)*



We discussed potential approaches to model the concept of "trust in automation" in the driving domain. In particular, we discussed an experimental design that could allow us to model trust as a function of user interventions and monitoring (i.e., "occlusion" of the driving environment) behavior when a driver cooperates with a lane-keeping system (i.e., an SAE level 2 driving automation system).

Assuming that both monitoring and intervention behavior are related to a users' subjective trust, the experiment could work the following way:

- A participant experiences an imperfect lane keeping system, i.e., the automation keeps the vehicle in the lane center but frequently starts to deviate from the ideal trajectory, slowly towards the lane boundaries.
- We assume that a driver observing this process would at some point start to correct the maneuvering of the vehicle to maintain a trajectory in the lane center again.

- While observing the situation and frequently correcting the steering, the participant needs to complete a (comparably short) trust questionnaire multiple times during the experiment.
- The collected data will then be analyzed to determine if the deviation that a driver allows the vehicle to leave the ideal trajectory is linked to the subjective trust ratings. We hypothesize, that a driver with low trust would correct the vehicles' maneuvers earlier and more frequently, than a driver who trusts the automation more.
- Additionally, we could include a secondary task to extend the principle to monitoring behavior. In other words, a high-trusting driver may monitor the driving automation system less often than a low-trusting driver.
- The data observed in both conditions (i.e., intervening/monitoring) could then be used to model the subjective trust of a driver. Given that the assumptions described above are valid, a drivers' monitoring and intervention behavior can then be used to derive their trust levels.

# 5 Panel discussions

## 5.1 Summary of Panel 1: How can models inform design?

*Antti Oulasvirta (Aalto University, FI), Alexandra Bremers (Cornell Tech – New York, US), Lewis Chuang (LMU München, DE), Debargha Dey (TU Eindhoven, NL), Andreas Riener (TH Ingolstadt, DE), and Shadan Sadeghian Borojeni (Universität Siegen, DE)*



Models are most useful if they are more than abstract, theoretical vehicles. They should not live in a vacuum, but be related to problems and issues in the real world. Therefore, the seminar wanted to discuss how models can inform the design of (in-)vehicle technology, and how they can inform policy. As both of these topics can fill an entire Dagstuhl by themselves, the primary objective was to identify the most pressing issues and opportunities.

**Andreas Riener** raised the question who we are designing for: the driver, the passenger, or both? The different stakeholders have varying demands for design. He then brought up the topic of cooperation with the OVID model of Roberts (1998).

**Shadan Sadeghian** followed up with the question on the blurring of boundaries between the vehicle and the human driver, which is further confused by the increasing levels of automation. She then discussed the dfiferent levels and their differing requirements posed to design.

**Dave Dey** discsused experimental research on automated vehicles and other road users, zooming into his own work on short-sighted aspects of interaction, like how pedestrians respond to a particular event. He raised the issue that research in this domain is complicated by multiple variables and it is hard to design solid research studies that can tease apart causes from confounds.

**Alexandra Bremers** took an interaction design point-of-view, talkign about how wizard of oz method can be used to study how people respond to automated vehicles, how grounding happens. However, this is confounded by the fact that automated vehicles are conceptually changing: they are not just transportation vehicles but becoming something something more: spaces for coming together, sharing with other people etc.

**Antti Oulasvirta** discussed design as counterfactual thinking. Consequently, the goal is to understand how design affects the adaptation of human behavior. He talked about emerging theories from ML, especially reinforcement learning, which offers a way to simulate the emergence of adaptive behavior.

**Lewis Chuang** raised the issue of what a model is, proposing a definition: "A representation of what might be going on "out there" based on our biological perception". This compares to information transer. Such models are important to avoid heuristic reasoning when thinking about automated driving. John Senders pioneering work is a great example of combining theoretical work with modeling and rigorous experimental research.

After the opening statements, the panel and the audience discussed three questions related to this theme:

**(1)** Types of questions: what types of questions exist at a design and policy level about human-automated vehicle interaction?

**(2)** How to inform decisions: How can models be used to inform design and policy decisions? What level of detail is needed here? What are examples of good practices?

**(3)** Integration: Integration can be considered in multiple ways. First, how can ideas from different disciplines be integrated (e.g., behavioral sciences, engineering, economics), even if they have at times opposing views (e.g., monetary gains versus accuracy and rigor)? Second, how can models become better integrated in the design and development process as tools to evaluate prototypes (instead of running empirical tests)? And third, how can models be integrated into the automation (e.g., as a user model) to broaden the automation functionality (e.g., prediction of possible driver actions, time needed to take over)?

As possible future topics, the participants saw three topics rise above others: 1) defining better what models do in design, especially what they predict; 2) finding connections between qualitative and quantitative understandings; and 3) making models more accessible for designers, i.e easier to use and better integrated to their practices.

## 5.2 Summary of Panel 2: What phenomena and driving scenarios need to be captured in computational models of human-automated vehicle interaction?

*Martin Baumann (Universität Ulm, DE), Luisa Heinrich (Universität Ulm, DE), Andrew Kun (University of New Hampshire – Durham, US), Dietrich Manstetten (Robert Bosch GmbH – Stuttgart, DE), Nikolas Martelaro (Carnegie Mellon University – Pittsburgh, US), and Hatice Sahin (Universität Oldenburg, DE)*



The second panel discussed what phenomena and driving scenarios need to be captured in computational models of human-automated vehicle interaction. The aim of this panel discussion was, as described by the organizers of the Dagstuhl Seminar, to both advance theory on human-automation interaction while also contributing to understanding realistic case studies for human-automation interaction that are faced for example by industry and governments. The following examples of possible relevant phenomena and driving scenarios were identified during the preparation of the seminar:

1. Transitions of control and dynamic attention: When semi-automated vehicles transfer control of the driving task back to the human, they require accurate estimates of a user's attention level and capability to take control (e.g., [14, 16])
2. Mental models, machine models, mode confusion, and training and skill: Models can be used to estimate human's understanding of the machine and vice-versa (e.g., [15]). Similarly, they might be used to estimate a human driver's skill level, and whether training is desired.
3. Shared control: In all these scenarios, there is some form of shared control. Shared control requires a mutual understanding of human and automation. Computational models can be used to provide such understanding for the automation (e.g., [17]).

These three areas and others were mentioned and discussed during the panel discussion. It started with short pitches by the panelists that are shortly summarized in the following section.

The following attendees were speakers on this panel: Luisa Heinrich, Nikolas Martelaro, Andrew Kun, Dietrich Manstetten, Hatice Sahin, Martin Baumann

**Luisa Heinrich** emphasized the importance of the take-over scenario as one key area where computational models of driver-automated vehicle interaction could be helpful. The main challenge here is how to design transitions strategies that support a safe and efficient transfer of control from the automation to the human driver, especially in cases when the human driver is out of the loop.

**Nikolas Martelaro** pointed out the importance of the complexity of traffic scenarios in which human drivers are required to take back control from the vehicle automation. A main target area for computational modelling could be how humans manage their situation awareness in such complex traffic situations given that there are many types of road users that have to be taken into account, such as pedestrians, cyclists, buses, passenger cars. He also emphasized that different kind of drivers need to be considered. Finally he mentioned as a last highly relevant phenomenon that automation might reduce some types of accidents but also leads to new kinds of accidents.

The third pitch was given by **Andrew Kun** and he focussed on the importance of bumper-to-bumper traffic as one relevant research area where computational models might be useful. The other topic Andrew brought into the discussion are the small bursts of interaction that occur during driving with automated vehicles where either the car or the human only contribute to the driving task for a short period of time. Additionally, the different types of tasks that are carried out while driving should be considered as relevant topic, such as non-driving related tasks, visual tasks, etc.

**Dietrich Manstetten** structured his pitch around for essential questions: What? Where? What for? In which situation? Regarding "what" Dietrich distinguished status vs. behavior models. The question here is what is the input and what is the output of a model. So the behavior could be used as input for status models of driver attention, whereas the current attention allocation could be an input to a computational model of lane change behavior. Regarding the "where" question he raised the points of lab vs. series vehicles. Whereas online restrictions are extremely relevant for series vehicles lab vehicles can work with offline data analysis tasks and have the freedom for highly sophisticated sensors that are nearly impossible in series vehicles. With regard to the "what for" question the main question is what is model used for and the main distinction is between safety and comfort. Which purpose is addressed determines the requirements on the quality of the computational models. Regarding the question "in which situation" Dietrich Manstetten opposed the take-over situation with the automation phase. He made clear that it is highly important to consider the environment and the situation any model was designed to work in. More aspects of the situation that should be considered are the traffic situation, the subtasks of the driving task, the time horizon or the vehicle type.

**Hatice Sahin** emphasized the fact that the general traffic ecosystem in which behavior occurs is generally neglected in studies and she raised the question how to integrate this complexity into computational models. Additionally, situational aspects such as traffic density and the presence of especially vulnerable road users need to be integrated in modelling activities.

The last pitch was presented by **Martin Baumann**. He focussed on the question of how humans interact with automated systems in traffic. He pointed out that the problem in addressing this question is in many cases not a lack of data, but a lack of precise definitions of the relevant constructs and predictive models. Computational models might be one promising way therefore to generate significant progress in answering this question. For this he mentioned three points that need to be addressed with computational models: The first is how people construct and maintain situational awareness in dynamic traffic situations as this mental representation is the basis for the decisions humans take. Second, long-term effects of the interaction with automated vehicles are highly important to consider. And third, more complex and realistic situations have to be considered in order to really evaluate the effect of automated systems on human behavior in traffic.

In the subsequent discussion the following topics, among others, where raised:

1. What kind of data can be used as a basis for models and especially how can data from the real world be integrated in the formation of computational models? There are big data sets, especially at companies that would be very helpful in building and validating models if they could be accessed by the scientific community.
2. With regard to the question which phenomena should be addressed with computational models the distinction was made between a technological perspective – what scenarios / environments should be addressed – and a human-centered perspective – what are the most important cognitive processes determining human behavior in traffic across different scenarios. Starting from the situation and the environmental context might allow to identify those cognitive phenomena that are relevant and that should be integrated into computational models of human behavior in these situations. On the other hand the point was made that it might be quite difficult to identify those scenarios that are really relevant in human-automated vehicle interaction as automated vehicles are still developing and the data we currently collect and use might not be relevant ones in the future. Additionally, starting from specific situations might lead to a situation where different theories and models for different situations are developed that are not compatible and cannot be generalized across different contexts. With regard to this the need for a unified and integrated modeling approach was formulated.
3. As one important situational factor the complexity of the situation was mentioned, and here it was stressed that it is not so much the "objective" complexity that is relevant, but the subjectively experienced complexity.
4. The topic of trust into technology and how it could be modelled computationally was discussed in depth. The different types of models and theories that the participants of the discussion use to investigate and measure trust were collected during the discussion. Relevant phenomena could be overtrust and distrust, uncertainty as a possible mechanism underlying trust.
5. One of the phenomena that was also mentioned in the context of the trust discussion was learning as a highly relevant phenomenon for computational models of human-automated

vehicle interaction. The experience with automated vehicles will shape the future behavior and currently many models of human-automated vehicle interaction are based on first time encounters with the technology.

The following papers where mentioned during the discussion and listed in the notes of the discussion:

### References


1 Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. The Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792

2 Teodorovicz, T., Kun, A. L., Sadun, R., & 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. https://doi.org/10.1016/j.ijhcs.2022.102789

3 Markkula, G., & Dogar, M. (2022). How accurate models of human behavior are needed for human-robot interaction? For automated driving?. arXiv preprint arXiv:2202.06123.

4 Sibi, S., Balters, S., Fu, E., Strack, E., Steinert, M, & Ju, Wendy. (2020). Back to School: Impact of Training on Driver Behavior and State in Autonomous Vehicles. 10.1109/IV47402.2020.9304537.

5 Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors, 53(5), 517–527. https://doi.org/10.1177/0018720811417254

6 Khavas, Z. R. (2021). A Review on Trust in Human-Robot Interaction. arXiv preprint arXiv:2105.10045.

7 Khavas, Z. R., Ahmadzadeh, S. R., & Robinette, P. (2020, November). Modeling trust in human-robot interaction: A survey. In International Conference on Social Robotics (pp. 529-541). Springer, Cham.

8 Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

9 Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. Journal of Business Research, 58(4), 500–507. https://doi.org/10.1016/S0148-2963(03)00140-1

10 Lee, J. G., & Lee, K. M. (2022). Polite speech strategies and their impact on drivers' trust in autonomous vehicles. Computers in Human Behavior, 127, 107015.

11 Jussi P. P. Jokinen and Tuomo Kujala. 2021. Modelling Drivers' Adaptation to Assistance Systems. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '21). Association for Computing Machinery, New York, NY, USA, 12–19. https://doi.org/10.1145/3409118.3475150

12 Kujala, T., & Lappi, O. (2021). Inattention and Uncertainty in the Predictive Brain. Frontiers in Neuroergonomics, 2, 718699. https://doi.org/10.3389/fnrgo.2021.718699

13 Azevedo-Sa, H., Jayaraman, S. K., Esterwood, C. T., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). Real-Time Estimation of Drivers' Trust in Automated Driving Systems. International Journal of Social Robotics, 13(8), 1911–1927. https://doi.org/10.1007/s12369-020-00694-1

14 Janssen, C. P., Iqbal, S. T., Kun, A. L., & Donker, S. F. (2019). Interrupted by my car? Implications of interruption and interleaving research for automated vehicles. International Journal of Human-Computer Studies, 130, 221-233.

**15** Janssen, C. P., Boyle, L. N., Kun, A. L., Ju, W., & Chuang, L. L. (2019). A hidden markov framework to capture human–machine interaction in automated vehicles. International Journal of Human–Computer Interaction, 35(11), 947-955.
**16** Wintersberger, P., Schartmüller, C., & Riener, A. (2019). Attentive User Interfaces to Improve Multitasking and Take-Over Performance in Automated Driving: The Auto-Net of Things. International Journal of Mobile Human Computer Interaction, 11(3), 40-58.
**17** Yan, F., Eilers, M., Weber, L., & Baumann, M. (2019). Investigating Initial Driver Intention on Overtaking on Rural Roads. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (pp. 4354-4359).

## 5.3 Summary of Panel 3: What technical capabilities do computational models need to possess?

*Antti Oulasvirta (Aalto University, FI), Jelmer Borst (University of Groningen, NL), Martin Fränzle (Universität Oldenburg, DE), Myounghoon Jeon (Virginia Polytechnic Institute – Blacksburg, US), Otto Lappi (University of Helsinki, FI), Gustav Markkula (University of Leeds, GB), and Nele Rußwinkel (TU Berlin, DE)*



The panel discussed technical capabilities required of computational models in this domain. Although the nature of different modeling frameworks and different studies might differ, what do we consider the core functionality? For example, there are capabilities related to:

**(1)** Compatibility: To what degree do models need to be compatible with simulator software (e.g., to test a "virtual participant"), hardware (e.g., be able to drive a car on a test track), and other models of human thinking?

**(2)** Adaptive nature: Computational models aim to strike a balance between precise predictions for more static environments and being able to handle open-ended dynamic environments (like everyday traffic). How can precision be guaranteed in static and dynamic environments? How can models adapt to changing circumstances?

**(3)** Speed of development and broader adoption: The development of computational models requires expertise and time. How can development speed be improved? How can communities benefit from each other's expertise?

The panel consisted of the following speakers: Gustav Markkula, Martin Fraenzle, Myounghoon Jeon (Philart), Jelmer Borst, Otto Lappi, and Nele Russwinkel.

**Gustav Markkula** discussed the compatibility of software and hardware, which is critical at the very end of an applied project, but much less considered earlier in the model developemnt stage. It is therefore critical to study the ultimate application context at least roughly. Similarly from an applied perspective, the speed of development and the flexibility of modelling is important. Models should be extensible with new data, new features, new capabilities etcetera.

**Martin Fraenzle** followed up and discussed how the idea of running models "in situ" might be out of scope at the moment.

**Myounghoon Jeon** brought up Distract-R as a great example of modeling, which allows different parameters to be simulated and rapid prototyping while being hardware and software agnostic and yet have some ecological validity. He echoed the need for adaptability and extensibiliy. He further brought up the topic of motions, which is presently missing from cognitive models although clearly important.

**Jelmer Borst** discussed the desiredata of modeling, one of which is that they should be able to drive the car. However, ACT-R is far from that goal according to the panelist. It is too brittle. There is a need for hybrid models that combine cognitive architectures and machine learning to overcome this issue. This may allow a cognitive model to operate in the background while the user is driving and warn users if something out of line is taking place, such as overloading.

**Otto Lappi** proposed a more philosophical perspective. He claimed that the first question to ask is who you are modeling for: the designer, the engineer, the scientist, or someone else? This has significant implications to the capabilities we want from models, and also our validation criteria.

**Nele Russwinkel** followed up on the topic of what is required of a model during driving, pointing out that ACT-R may be useful if it can anticipate the driver, even if now able to drive the car. She pointed out that latent variables, such as those related to situation awareness, are important for a car to understand. However, presentday models mostly do not afford this in a real-time system, due to computational intensity. She pointed out that a lot of modeling is still missing when it comes to the different levels of automation in vehicles. Moreover, some of the events we are interested in are rare and therefore inherently difficult to predict.

Discussion with the audience followed up on the need for better simulation environments and hybrid models that use ML. There was disagreement on whether simulation environments are too simplistic and can or cannot contain the richness of perceptual and other cues that drivers exploit in driving. The audience gravitated toward the need for developing a roadmap for cognitive models. Participants agreed that one of the outstanding goals is to understand the different SAE levels better, especially from the perspective of their psychological consequences and, therefore, their consequences on cognitive models.

Future topics that were raised included:

1) How to allow drivers update models about themselves?
2) How to trade off predictive accuracy and computational costs in such a way that we can allow cognitive models to run in real-time systems?
3) How to go from mental models research to scene understanding, which is critical in driving.

## 5.4 Summary of Panel 4: How can models benefit from advances in AI while avoiding its pitfalls?

*Christian P. Janssen (Utrecht University, NL), Duncan Brumby (University College London, GB), Birsen Donmez (University of Toronto, CA), Justin Edwards (ADAPT Centre – Dublin, IE), Mark Eilers (Humatects – Oldenburg, DE), Moritz Held (Universität Oldenburg, DE), Jussi Jokinen (University of Jyväskylä, FI), and Roderick Murray-Smith (University of Glasgow, GB)*



The fourth panel discussed how computational models that are developed for Human-Automated Vehicle Interaction can benefit from advances in AI while also avoiding some of its pitfalls. The summary below is made by Chris Janssen, building on crowd-sourced community notes of the sessions and snippets of panel members' presentations or text where possible.

Before the meeting, the organizers had identified that there are many developments in AI that computational models can benefit from. Three examples are advances in (1) simulator-based inference (e.g., [1]) to reason about possible future worlds (e.g., varieties of traffic environments), (2) reinforcement learning [2] and its application to robotics [3] and human driving [4], and (3) deep learning [5] and its potential to predict driver state or behavior from sensor data. At the same time, incorporation of AI techniques also comes with challenges that need to be addressed. Three potential challenges are for example:

1. Explainability: Machine learning techniques are good at classifying data, but do not always provide insight into why classifications are made. This limits their explainability and is at odds with the objective of computational models to gain insight into human behavior. How can algorithms' explainability be improved?
2. Scalability and generalization: How can models be made that are scalable to other domains and that are not overtrained on specific instances? How can they account for future scenarios where human behavior might be hard to predict [6]?
3. System training and corrective feedback: if models are trained on a dataset, what is the right level of feedback to correct an incorrect action to the model? How can important new instances and examples be given more weight to update the model's understanding without biasing the impact?

During the panel, these and other themes were discussed. We started with short pitches by the panelists, which are summarized first.

**Roderick Murray-Smith** discussed among others the following points.

**(1)** A priori insight can be placed into machine learning models in multiple ways. Forward and inverse modelling components were discussed as being particularly relevant.
**(2)** The advent of deep learning models has helped to place more complex models of perception into control models.
**(3)** This has knock-on effects of explainability: is trying to explain a complex policy model a good idea? Or is it better to ensure that the model learns a value function that is relevant to the human and then optimized?
**(4)** Engineering approaches often follow a modular approach, where multiple components are solved from first-principles, before being placed in bigger systems. The question is whether such an approach is scalable to a more complex context such as driving, where a closed -loop simulation might not be feasible.

**Mark Eilers** discussed that the best models for driving an automated vehicle do not necessarily need to be models that drive in a human way, or based on a human-like model. Mark saw potentials and possibilities in the following fields and areas.

**(1)** robotics, for example imitation learning,
**(2)** reinforcement learning and optimal control,
**(3)** deep learning, for example the use of new unsupervised learning techniques.

All these techniques can be adapted to fit existing models and contexts. Mark identified as pitfalls that there might be an identifiability problem: data alone can't always tell the researcher which of the many different approaches or models is the correct one. Theory and experiments together must inform the model and the modeling community.

**Duncan Brumby** reflected on his own past work in preparation for the meeting. When he modeled driving 15 years ago, he used a relatively simple model, but it was useful and valuable (e.g., [7]). At the time, there was talk of how reinforcement learning models can one day give even more insight, and it is good to see that that potential has come to fruition (e.g., [4]). Duncan now sees three areas where science of AI has lead to great developmens and still has potential to grow further.

**(1)** Models of the driver: science can help to create better models
**(2)** Models for the car itself: current semi-automated vehicles have lots of technology and models in it, for example to go around a corner and break when a car in front of it stops. However, the models for these tasks are developed by companies, and the question is how it will reach science.
**(3)** The technology with which people interact in the car (e.g., secondary devices) have evolved. The tools that people use are different.

Overall, due to these changes, some of the parameters that underly older models might need updating. For example, when do we call something distracting or not differs between settings.

**Birsen Donmez** looked at the topic of this panel from a statistics and experimentation perspective. Some machine learning approaches such as Deep Learning are very powerful, as they enable to model complex non-linear systems. Capturing multiple predictors and their interactions go beyond the model typical traditional statistical or cognitive models. However, they do require big data sets or samples. More traditional statistical tools do not always require such larger sets, but can also work on smaller samples. There are challenges for both the big data set (and machine learning) and small data set (and statistics) approaches. For the bigger data sets the general challenge is that these are often collected by industry and not available for researchers. How can this be changed? When we collect our own datasets they might also not always be big enough, and when using data sets of others, that might also not always be trivial. For smaller data sets there is a limited sample problem, when applying machine learning techniques. It requires insight to for example split the dataset appropriately into training set and test set. Birsen sees potential for situations where researchers do get hold of bigger data sets. They have the right background to be critical about the data. For example, to ask questions such as: is the data representative? How was data collected? Do correlations exist?

**Jussi Jokinen** discussed a model of how humans can interact with an automated vehicle. A cooperative AI can for example observe human user behavior, and then try to identify posterior probabilities of user states (in a Bayesian fashion). There are however potential pitfalls in that such a Bayesian inference process can be slow and the results of a modeling framework are sometimes surprising and don't seem plausible. Given that the model is not always correct, it is useful to explicitly consider (un-)certainty. In Jussi's own work (e.g., [4]), he benefits from applying the computational rationality framework (e.g., [8, 9]). Such a model expresses how latent variables impact behavior. A reinforcement learning agent is developed that works within constraints of a cognitive architecture and task environment to come up with the optimal policy that if architecture and task are correctly defined should be similar to user's behavior. It is cool when it works, but there are lots of moving parts. Jussi sees opportunities (and challenges) in refinement of the ability to use inference techniques in these models: how can one be confident that the learned model is indeed the correct model?

**Justin Edwards** drew parallels with the history of language models and his experience in the Conversational User Interfaces (CUI) community. Coming up with language generation and processing models long proved a challenge to the field. Initially there were many rule-based models. The last decade has seen a wider range of models that use large datasets and are data-driven. Data-driven models have surpassed the more theory driven models in quite a lot of tasks. However, they also are occasionally deeply flawed. The samples on which training takes place are not always general, but are used for generalization purposes. This showed for example in the Delphi system, which was meant to be capable of moral judgements, but used data from reddit chatboards for training.So, the big lesson learned for the automotive community is: as new technology becomes available, be aware of where the data comes from. What are the biases? Those might carry through in surprising ways.

**Moritz Held** focused mostly on two themes. First: explainability. Hybrid models that combine machine learning with theory-driven models might be fruitful in this regard. For example, combining ACT-R models with techniques from Bayesian spatial networks. The second topic was scalability of models. Moritz was surprised by some discussions in previous days regarding scalability. He would expect that there is always some robustness in a model. There is no need for (for example) and ACT-R supermodel to handle every driving situation. But, he would expect it to at least handle different driving scenarios.

**Chris Janssen** talked about four points. First, it is important to think about function allocation: how are tasks divided between humans and the automated vehicles. Models can play a role in this regard (see e.g. [10]). Second, in such discussions, it is tempting to think of simple heuristics such as "men are better at [some tasks] and machines are better at [other tasks]" [11]. However, humans and machines are dynamic (they can both learn and adapt), and they operate in varying contexts. Therefore, static heuristics might be useful at first, but might not always be appropriate for these dynamic contexts. Cognitive models should take such dynamics into account. Third, explainability is becoming increasingly important inside and outside of science. Computational models can tie in with this by combining theory-driven and data-driven models. In other words: it is important that models are connect to theory (not just data), but also that theory is connected to practice.

During the discussion that followed, we talked among others about the following aspects:

1. What creates distributional shifts in models? How can models be made more robust against this?
2. How can the designers and researchers of modelers better communicate what the goals and limitations of models are? For example, for which level of automation they are designed.
3. There seems to be a balance between generalization of a model versus having a more structured / fixed model. Models differ in which aspects they keep fixed (as assumptions, framework, or architecture) and which aspects they train or learn (for example, based on data).
4. There was also a discussion about how computational models can be designed such that designers and developers (of cars) can better benefit from them. There is a potential discrepancy here in that designers want to test multiple, flexible designs whereas models are often trained on or designed for more fixed tasks.
5. There is a need to be more systematic about the model selection process.
6. There is an open question about for what types of research questions small or large datasets are needed.

### References


1 Kangasrääsiö, A., Jokinen, J. P., Oulasvirta, A., Howes, A., & Kaski, S. (2019). Parameter inference for computational cognitive models with Approximate Bayesian Computation. *Cognitive Science*, 43(6)

2 Sutton, R., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press

3 Levine, S. (2018). *Reinforcement learning and control as probabilistic inference: Tutorial and review.* arXiv preprint arXiv:1805.00909.

4 Jokinen, J. P., Kujala, T., & Oulasvirta, A. (2021). Multitasking in driving as optimal adaptation under uncertainty. *Human factors*, 63(8), 1324-1341.

5 Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* Cambridge, MA: MIT press.

6 Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man–machine systems* (pp. 129-135). Pergamon.
7 Brumby, D. P., Salvucci, D. D., & Howes, A. (2009, April). Focus on driving: How cognitive constraints shape the adaptation of strategy when dialing while driving. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1629-1638).
8 Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6(2), 279-311.
9 Oulasvirta, A., Kristensson, P. O., Bi, X., & Howes, A. (Eds.). (2018). *Computational interaction*. Oxford University Press.
10 Janssen, C. P., Boyle, L. N., Kun, A. L., Ju, W., & Chuang, L. L. (2019). A hidden markov framework to capture human–machine interaction in automated vehicles. *International Journal of Human–Computer Interaction*, 35(11), 947-955.
11 Fitts PM (ed) (1951) *Human engineering for an effective air navigation and traffic control system*. National Research Council, Washington, DC

## 5.5 Summary of Panel 5: What insights are needed for or from empirical research?

*Shamsi Tamara Iqbal (Microsoft – Redmond, US), Linda Ng Boyle (University of Washington – Seattle, US), Benjamin Cowan (University College – Dublin, IE), Patrick Ebel (Universität Köln, DE), Wendy Ju (Cornell Tech – New York, US), Tuomo Kujala (University of Jyväskylä, FI), Philipp Wintersberger (TU Wien, AT), and Fei Yan (Universität Ulm, DE)*



Computation models are only as good to the degree that they can describe and predict phenomena in the real world. In particular, many current computational models capture the results of a single experiment based on the current context of what is being modeled. However, behavior might change with more exposure to and experience with automated technology. To make models most useful over time, empirical testing is of paramount importance, especially in order to evaluate models that look at current behavior, as well as models that are based on projections of future behavior. The panel on empirical research insights focused on the question above, as well as related questions such as how to study phenomena where computational models are not commercially available, and how to map findings from simulator tests to real world scenarios.

**Patrick Ebel** started the conversation with a focus on mapping real world data to input for computational models. Clearly there is a need for models to be used in industry and leveraging existing models can save time and money but there seems to be a barrier in adaptation. A couple of possible causes could be accuracy of models, especially since real world application needs to consider many confounding factors which cannot be typically captured in experimental models. Patrick also raised the question of the comparability of data that come from different simulators and how to reconcile for those differences.

**Wendy Ju** discussed the generalizability of empirical studies that are done in controlled settings in the lab – where the studies themselves are challenging to run and have additional limitations (e.g. cultural, temporal variation). She proposed the alternative of gleaning results from understanding naturalistic data from real-world driving. However, this will require instrumentation of vehicles where there needs to be a common agreement of what data is collected, when and how. The data itself can be a contribution to the field, and once we have these datasets, ML can be used to scale up and detect patterns in the diverse data sets that are collected.

**Fei Yan** talked about how to extract insights from empirical research and black box models and posed the core question of determining when a model is really needed to solve a research question. Empirical research can help understand causal relationships, and black box models can look at theory (e.g. social psychology on human-human interaction) which can be then used to form hypotheses that can be tested in empirical research. Fei proposed using empirical research to understand relevant factors before making the black box model. An example scenario could be the use of a predictive model in a lane change support system that can adapt to a driver's states of uncertainty. But it is difficult to know what models are most applicable in these scenarios. For some domains, ML or Bayesian models might be enough, but depending on the level of accuracy needed, those might not be most appropriate for other domains. Fei also talked about the small sample size in empirical research which helps with deeper dives into relationships between factors.

**Tuomo Kujala** shifted gears towards modelling driver attention. Driver attention monitoring is required for cars in level 3 or even up to level 4 in order for a car to respond safely and effectively to takeover requests. The challenge is that most attention systems monitor the inside of the car and how the driver's attention is impacted by the internal environment that can be detected. However, a driver's attention can be impacted by factors that are not detectable directly – for example, how much 'free attention' or cognitive capacity is available for the driver and whether it is enough to handle an upcoming scenario in driving. Appropriate attention modelling can decrease the uncertainty. Tuomo proposed that we need prescriptive or normative computational models, not only descriptive models.

**Phillipp Wintersberger** focused on the specific scenario of what is needed to make Level 3 automated driving possible and presented the argument that simulator results may not be as limited as we assume them to be. The trade off for a data-driven model is that a huge amount of naturalistic data is needed, connecting back to Wendy's point – we as a field do not have established standards on how to get this data. For example, doing a study in a test track may not yield useful data if the test track is not similar to the real world, but similar studies can be designed to be more similar to real world scenarios. Phillipp also talked about how we can get massive data to investigate development of trust in such systems. Is gamification a possible path? He also questioned the need for really big data sets and whether smaller datasets can be combined to get bigger sets.

**Shamsi Iqbal** talked about the target user – who we are designing for, the current user or the future user who will have different technological capabilities, exposure and experience. Can we effectively extrapolate from the current user to project what the future user will look like? Empirical research typically falls into three buckets – lab experiments, model building approach and naturalistic driving. While naturalistic driving gives the opportunity to learn more about what challenges people face in real world scenarios, there are also privacy and security concerns around collecting massive amounts of data from people. For lab experiments in the other hand, we have the opportunity to test innovative ideas that push the boundary of what might be possible. The challenge is then to determine how to generalize the findings to real world scenarios, where we may not even know what the future scenarios might be.

**Linda Boyle** brought up the counter point that model-based and empirical research are not mutually exclusive and naturalistic and simulated data should not be mutually exclusive either. Research needs data from separate dimensions, not just comparative studies. Even in a single lab – different types of data can be collected. The goal should not be just to create a data repository, but also to use results for scenarios such as Bayesian inferences/priors for future models and experiments. Linda emphasized that while field studies are great there is a lot of variation. Lab studies are a must – especially to test rate situations and edge cases that we may rarely or never see in real world data, but still need to be accounted for in our models.

**Ben Cowan** emphasized that the key to empirical research is observation of phenomena. In his view, the role of empirical research in automated vehicles are: a) to assess how concepts we deem important to automated driving (e.g. situational awareness, attention) manifest in and are impacted within the experience of automated driving. We can learn a lot from conducting qualitative studies on automated driver experience (e.g. observation studies, interviews) to assess the importance of concepts like trust, potential challenges and barriers to adoption etc. Quantitative empirical work, particular lab studies can also be used to test how specific concepts are causally influenced by design or events during the drive to get a clearer picture of what are the key areas for us to focus on (e.g. the design of communication strategies for pedestrians- Lanzer et al, AutoUI 2020 on politeness strategies positive effect on trust and acceptance). They also allow us to experiment with identifying cognitive concepts of importance.

To support the notion of what to model and to give our initial models a test run: Models should be related to the real world. But we need to perhaps incubate these first. Empirical lab based work is important to determine what is and is not important in that real world for the concepts we deem important to research. That is not to say that we cannot and should not do lab-based work, devise, test and assess models for concepts on lab based data to determine how the cognitive concepts may behave within more real world driving contexts.

A key consideration is also in the measurement quality for measuring the concepts we wish to measure. In particular when using questionnaires to assess concepts like trust, it is important that we stay true to any base questionnaires used or conduct the design of our own psychometrics to ensure what we are using are valid, reliable and sensitive to the concepts we measure. On a wider point as a community of researchers we must also ensure that the empirical work we are doing is replicable and open. We need to create an infrastructure and reward system that allows us to embed replication in our empirical research activities (either through replication when publishing the initial study, or in specific tracks at our key conferences and journal publication venues).

In the discussion, Patrick talked about triangulating empirical research and model based research using the following approach: 1) Use model based systems that are built upon naturalistic data. Can help to see patterns in the data and suggest empirical tests, and 2) Use models as a sandbox and test bed approach. How good does a model need to be? If model already shows a study is not worth the time, might be useful to test. Another important topic that emerged from the panel discussion is whether a 'replicate' track should be added to the Auto-UI conference. Key questions that arose included what is the definition of replication, what is the expected rigor, how to address the differences in the experimental setup that might impact the replication, how to share data, how to present the results at the conference and what kind of benchmarks should be setup.

As possible future topics, the following points were extracted from the panel discussion:

- How to connect models even better to needs of industry? They want them, but don't always use them. Why? (e.g., Patrick)
- How to move towards "data is a contribution" and replication (Wendy)
- Do we need models even? (e.g., Fei)

- Prescriptive models are needed; not just descriptive (Tuomo)
- To understand attention, more is needed than just in-car glances (Tuomo)
- Do we need realistic studies? Or is simulator sometimes better (Philipp)
- Opportunity of gamification (Philipp)
- Need for a place to store online data (Philipp)
- Who are we designing tech / experiment / model for? Today's user or tomorrow's user? (Shamsi)
- How to make tools accessible to wider community? For models and tools and experiments. (Linda)
- How to motivate and reward replication systems? (Ben)

# 6 Open problems

## 6.1 Relevant papers for modeling human-automated vehicle interaction

*Christian P. Janssen (Utrecht University, NL)*



Among the attendees we also gathered an overview of papers that they thought were interesting for researchers in the field.

Before the conference, we sent out a google form and asked attendees to submit papers that they thought were interesting and either written by themselves or others. The suggested papers are for example relevant domains for modeling, or examples of modeling papers. The following papers were suggested: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46]

During the conference, we also crowdsourced a collection of relevant papers. The large majority of these papers contain examples of models or (conceptual) frameworks or datasets that are used for or inspired by models. These papers were suggested: [3, 13, 15, 16, 37, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76]

### References

1 Anderson, J. R., Zhang, Q., Borst, J. P., & Walsh, M. M. (2016). The discovery of processing stages: Extension of Sternberg's method. Psychological review, 123(5), 481.

2 Bianchi Piccinini, G., Lehtonen, E., Forcolin, F., Engström, J., Albers, D., Markkula, G., Lodin, J., & Sandin, J. (2020). How do drivers respond to silent automation failures? Driving simulator study and comparison of computational driver braking models. Human factors, 62(7), 1212-1229.

3 Blum, S., Klaproth, O., & Russwinkel, N. (2022). Cognitive Modeling of Anticipation: Unsupervised Learning and Symbolic Modeling of Pilots' Mental Representations. Topics in Cognitive Science, 10.1111/tops.12594. Advance online publication. https://doi.org/10.1111/tops.12594

4 Cummings, M. L., Li, S., & Zhu, H. (2022). Modeling operator self-assessment in human-autonomy teaming settings. International Journal of Human-Computer Studies, 157, 102729.

5 Damm, W., Fränzle, M., Lüdtke, A., Rieger, J. W., Trende, A., & Unni, A. (2019, June). Integrating neurophysiological sensors and driver models for safe and performant automated vehicle control in mixed traffic. In 2019 IEEE Intelligent Vehicles Symposium (IV) (pp. 82-89). IEEE.

6 Eilers, M., Fathiazar, E., Suck, S., Twumasi, D. (2019) Dynamic Bayesian networks for driver-intention recognition based on the traffic situation. In Cooperative Intelligent Transport Systems: Towards high-level automated driving, , pp. 465-495

7 Funk Drechsler, M., Peintner, J. B., Seifert, G., Huber, W., & Riener, A. (2021, September). Mixed Reality Environment for Testing Automated Vehicle and Pedestrian Interaction. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 229-232).

8 van der Heiden, R. M., Kenemans, J. L., Donker, S. F., & Janssen, C. P. (2021). The effect of cognitive load on auditory susceptibility during automated driving. Human factors, 0018720821998850.

9 Janssen, C. P., Boyle, L. N., Ju, W., Riener, A., & Alvarez, I. (2020). Agents, environments, scenarios: A framework for examining models and simulations of human-vehicle interaction. Transportation research interdisciplinary perspectives, 8, 100214.

10 Janssen, C. P., Iqbal, S. T., Kun, A. L., & Donker, S. F. (2019). Interrupted by my car? Implications of interruption and interleaving research for automated vehicles. International Journal of Human-Computer Studies, 130, 221-233.

11 Jeon, M., Zhang, Y., Jeong, H., P. Janssen, C., & Bao, S. (2021, September). Computational Modeling of Driving Behaviors: Challenges and Approaches. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 160-163).

12 Jokinen, J. P., & Kujala, T. (2021). Modelling Drivers' Adaptation to Assistance Systems. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 12-19).

13 Jokinen, J. P., Kujala, T., & Oulasvirta, A. (2021). Multitasking in driving as optimal adaptation under uncertainty. Human factors, 63(8), 1324-1341.

14 Kambhampati, S. (2020). Challenges of human-aware AI systems. AI Magazine, 41(3), 3-17. https://doi.org/10.1609/AIMAG.V41I3.5257

15 Kanaan, D., Ayas, S., Donmez, B., Risteska, M., & Chakraborty, J. (2019). Using naturalistic vehicle-based data to predict distraction and environmental demand. International Journal of Mobile Human Computer Interaction (IJMHCI), 11(3), 59-70.

16 Klaproth, O.W., Halbrügge, M., Krol, L.R., Vernaleken, C., Zander, T.O., & Russwinkel, N. (2020), A Neuroadaptive Cognitive Model for Dealing with Uncertainty in Tracing Pilots' Cognitive State. Topics in Cognitive Science, 12(3), 1012-1029. doi:10.1111/tops.12515

17 Ko, S., Kutchek, K., Zhang, Y., & Jeon, M. (2022). Effects of non-speech auditory cues on control transition behaviors in semi-automated vehicles: Empirical study, modeling, and validation. International Journal of Human–Computer Interaction, 38(2), 185-200.

18 Krishnan, R., & Tickoo, O. (2020). Improving model calibration with accuracy versus uncertainty optimization. Advances in Neural Information Processing Systems, 33, 18237-18248.

19 Kuen, J., Schartmüller, C., & Wintersberger, P. (2021, September). The TOR Agent: Optimizing Driver Take-Over with Reinforcement Learning. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 47-52).

20 Lavin, A., Zenil, H., Paige, B., Krakauer, D., Gottschlich, J., Mattson, T., ... & Pfeffer, A. (2021). Simulation Intelligence: Towards a New Generation of Scientific Methods. arXiv preprint arXiv:2112.03235.

21 Lechner, M., Hasani, R., Amini, A., Henzinger, T. A., Rus, D., & Grosu, R. (2020). Neural circuit policies enabling auditable autonomy. Nature Machine Intelligence, 2(10), 642-652.

22 Lee, J. Y., & Lee, J. D. (2019). Modeling microstructure of drivers; task switching behavior. International Journal of Human-Computer Studies, 125, 104-117.

23 Leon, J. F. (2020, July). Robust Scene Understanding via Real-Time Approximate Bayesian Computations. In 2020 International Conference on Systems, Signals and Image Processing (IWSSIP) (pp. 13-13). IEEE.
24 Manstetten, D., Beruscha, F., Bieg, H. J., Kobiela, F., Korthauer, A., Krautter, W., & Marberger, C. (2020, October). The Evolution of Driver Monitoring Systems: A Shortened Story on Past, Current and Future Approaches How Cars Acquire Knowledge About the Driver's State. In 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 1-6).
25 Markkula, G., & Dogar, M. (2022). How accurate models of human behavior are needed for human-robot interaction? For automated driving?. arXiv preprint arXiv:2202.06123.
26 Martelaro, N., Teevan, J., & Iqbal, S. T. (2019, May). An exploration of speech-based productivity support in the car. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
27 Martín, J. A. Á., Gollee, H., Müller, J., & Murray-Smith, R. (2021). Intermittent control as a model of mouse movements. ACM Transactions on Computer-Human Interaction (TOCHI), 28(5), 1-46.
28 Morando, A., Gershon, P., Mehler, B., & Reimer, B. (2021). A model for naturalistic glance behavior around Tesla Autopilot disengagements. Accident Analysis & Prevention, 161, 106348.
29 Oulasvirta, A. (2019). It's time to rediscover HCI models. Interactions, 26(4), 52-56.
30 Oulasvirta, A., Kristensson, P. O., Bi, X., & Howes, A. (Eds.). (2018). Computational interaction. Oxford University Press.
31 Pakdamanian, E., Sheng, S., Baee, S., Heo, S., Kraus, S., & Feng, L. (2021, May). Deeptake: Prediction of driver takeover behavior using multimodal data. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-14).
32 Pakdamanian, E., Sheng, S., Baee, S., Heo, S., Kraus, S., & Feng, L. (2021, May). Deeptake: Prediction of driver takeover behavior using multimodal data. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-14).
33 Petermeijer, S. M., Tinga, A., Jansen, R., de Reus, A., & van Waterschoot, B. (2021, September). What Makes a Good Team?-Towards the Assessment of Driver-Vehicle Cooperation. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 99-108).
34 Rehman, U., Cao, S., & MacGregor, C. (2019). Using an integrated cognitive architecture to model the effect of environmental complexity on drivers' situation awareness. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 63, No. 1, pp. 812-816). Sage CA: Los Angeles, CA: SAGE Publications.
35 Ringfort-Felner, R., Laschke, M., Sadeghian, S., & Hassenzahl, M. (2022). Kiro: A Design Fiction to Explore Social Conversation with Voice Assistants. Proceedings of the ACM on Human-Computer Interaction, 6(GROUP), 1-21.
36 Sadeghian, S., Hassenzahl, M., & Eckoldt, K. (2020, September). An exploration of prosocial aspects of communication cues between automated vehicles and pedestrians. In 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 205-211).
37 Scharfe, M. & Russwinkel, N. (2019). Towards a Cognitive Model of the Takeover in Highly Automated Driving for the Improvement of Human Machine Interaction. In Proceedings of the 17th International Conference on Cognitive Modelling, Montreal, Canada (pp. 210-215).
38 Suo, S., Regalado, S., Casas, S., & Urtasun, R. (2021). Trafficsim: Learning to simulate realistic multi-agent behaviors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10400-10409).

**39** Teodorovicz, T., Kun, A. L., Sadun, R., & 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.

**40** Todi, K., Bailly, G., Leiva, L., & Oulasvirta, A. (2021, May). Adapting user interfaces with model-based reinforcement learning. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

**41** Tonolini, F., Radford, J., Turpin, A., Faccio, D., & Murray-Smith, R. (2020). Variational inference for computational imaging inverse problems. Journal of Machine Learning Research, 21(179), 1-46.

**42** Wintersberger, P., Schartmüller, C., Shadeghian-Borojeni, S., Frison, A. K., & Riener, A. (2021). Evaluation of imminent take-over requests with real automation on a test track. Human factors, 00187208211051435.

**43** Wong, P. N., Brumby, D. P., Babu, H. V. R., & Kobayashi, K. (2019). Voices in Self-Driving Cars Should be Assertive to More Quickly Grab a Distracted Driver's Attention. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 165-176).

**44** Wu, T., Martelaro, N., Stent, S., Ortiz, J., & Ju, W. (2021). Learning When Agents Can Talk to Drivers Using the INAGT Dataset and Multisensor Fusion. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(3), 1-28.

**45** Yan, F., Eilers, M., Lüdtke, A., & Baumann, M. (2016, June). Developing a model of driver's uncertainty in lane change situations for trustworthy lane change decision aid systems. In 2016 IEEE Intelligent Vehicles Symposium (IV) (pp. 406-411). IEEE.

**46** Zhang, Y., Wu, C., Qiao, C., Sadek, A., & Hulme, K. F. (2022). A Cognitive Computational Model of Driver Warning Response Performance in Connected Vehicle Systems. IEEE Transactions on Intelligent Transportation Systems.

**47** Abbink, D. A., Mulder, M., & Boer, E. R. (2012). Haptic shared control: smoothly shifting control authority?. Cognition, Technology & Work, 14(1), 19-28.

**48** Bärgman, J., Boda, C. N., & Dozza, M. (2017). Counterfactual simulations applied to SHRP2 crashes: The effect of driver behavior models on safety benefit estimations of intelligent safety systems. Accident Analysis & Prevention, 102, 165-180.

**49** Bianchi Piccinini, G., Lehtonen, E., Forcolin, F., Engström, J., Albers, D., Markkula, G., ... & Sandin, J. (2020). How do drivers respond to silent automation failures? Driving simulator study and comparison of computational driver braking models. Human factors, 62(7), 1212-1229.

**50** Colombi, J. M., Miller, M. E., Schneider, M., McGrogan, M. J., Long, C. D. S., & Plaga, J. (2012). Predictive mental workload modeling for semiautonomous system design: Implications for systems of systems. Systems Engineering, 15(4), 448-460.

**51** Gunzelmann, G., Moore Jr, L. R., Salvucci, D. D., & Gluck, K. A. (2011). Sleep loss and driver performance: Quantitative predictions with zero free parameters. Cognitive Systems Research, 12(2), 154-163.

**52** He, D., Wang, Z., Khalil, E. B., Donmez, B., Qiao, G., & Kumar, S. (in press). Classification of driver cognitive load: Exploring the benefits of fusing eye-tracking and physiological measures. Transportation Research Record.

**53** Janssen, C. P., & Brumby, D. P. (2010). Strategic adaptation to performance objectives in a dual-task setting. Cognitive science, 34(8), 1548-1560.

**54** Janssen, C. P., Brumby, D. P., & Garnett, R. (2012). Natural break points: The influence of priorities and cognitive and motor cues on dual-task interleaving. Journal of Cognitive Engineering and Decision Making, 6(1), 5-29.

55 Janssen, C. P., Boyle, L. N., Kun, A. L., Ju, W., & Chuang, L. L. (2019). A hidden markov framework to capture human–machine interaction in automated vehicles. International Journal of Human–Computer Interaction, 35(11), 947-955.

56 Janssen, C. P., Boyle, L. N., Ju, W., Riener, A., & Alvarez, I. (2020). Agents, environments, scenarios: A framework for examining models and simulations of human-vehicle interaction. Transportation research interdisciplinary perspectives, 8, 100214.

57 Janssen, C. P., Everaert, E., Hendriksen, H. M., Mensing, G. L., Tigchelaar, L. J., & Nunner, H. (2019). The influence of rewards on (sub-) optimal interleaving. PloS one, 14(3), e0214027.

58 Janssen, C. P., Iqbal, S. T., Kun, A. L., & Donker, S. F. (2019). Interrupted by my car? Implications of interruption and interleaving research for automated vehicles. International Journal of Human-Computer Studies, 130, 221-233.

59 Johora, F. T., & Müller, J. P. (2018, November). Modeling interactions of multimodal road users in shared spaces. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (pp. 3568-3574). IEEE.

60 Kahl, S., Wiese, S., Russwinkel, N., & Kopp, S. (2021). Towards autonomous artificial agents with an active self: modelling sense of control in situated action. Cognitive Systems Research, 72, 50-62. https://doi.org/10.1016/j.cogsys.2021.11.005

61 Khosroshahi, E. B., Salvucci, D. D., Veksler, B. Z., & Gunzelmann, G. (2016). Capturing the effects of moderate fatigue on driver performance. In Proceedings of the 14th International Conference on Cognitive Modeling (pp. 163-168).

62 Lotz, A., Russwinkel, N., Wagner, T. and Wohlfarth, E. (2020). An adaptive assistance system for subjective critical driving simulation: understanding the relationship between subjective and objective complexity. In D. de Waard et al. (Eds.), Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2019 Annual Conference. Nantes, France, p. 97-108.

63 Lotz, A., Wiese, S. & Russwinkel, N. (2019). SEEV-VM: ACT-R Visual Module based on SEEV theory. In Proceedings of the 17th International Conference on Cognitive Modelling (ICCM 2019), Montreal, Canada (pp. 301-306).

64 Markkula, G., Romano, R., Madigan, R., Fox, C. W., Giles, O. T., & Merat, N. (2018). Models of human decision-making as tools for estimating and optimizing impacts of vehicle automation. Transportation research record, 2672(37), 153-163.

65 Mole, C., Pekkanen, J., Sheppard, W., Louw, T., Romano, R., Merat, N., ... & Wilkie, R. (2020). Predicting takeover response to silent automated vehicle failures. Plos one, 15(11), e0242825.

66 Nagaraju, D., Ansah, A., Ch, N. A. N., Mills, C., Janssen, C. P., Shaer, O., & Kun, A. L. (2021). How Will Drivers Take Back Control in Automated Vehicles? A Driving Simulator Test of an Interleaving Framework. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 20-27).

67 Pekkanen, J., Giles, O. T., Lee, Y. M., Madigan, R., Daimon, T., Merat, N., & Markkula, G. (2021). Variable-drift diffusion models of pedestrian road-crossing decisions. Computational Brain & Behavior, 1-21.

68 Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. Human factors, 48(2), 362-380.

69 Salvucci, D. D., & Gray, R. (2004). A two-point visual control model of steering. Perception, 33(10), 1233-1248.

70 Salvucci, D. D., Zuber, M., Beregovaia, E., & Markley, D. (2005). Distract-R: Rapid prototyping and evaluation of in-vehicle interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 581-589).

**71** Schwall, M., Daniel, T., Victor, T., Favaro, F., & Hohnhold, H. (2020). Waymo public road safety performance data. arXiv preprint arXiv:2011.00038.
**72** Svärd, M., Markkula, G., Bärgman, J., & Victor, T. (2021). Computational modeling of driver pre-crash brake response, with and without off-road glances: Parameterization using real-world crashes and near-crashes. Accident Analysis & Prevention, 163, 106433.
**73** Trende A., Unni A., Rieger J., Fraenzle M. (2021) Modelling Turning Intention in Unsignalized Intersections with Bayesian Networks. In: Stephanidis C., Antona M., Ntoa S. (eds) HCI International 2021 – Posters. HCII 2021. Communications in Computer and Information Science, vol 1421. Springer, Cham.
**74** van Maanen, L., Heiden, R. V. D., Bootsma, S., & Janssen, C. P. (2021). Identifiability and Specificity of the Two-Point Visual Control Model of Steering. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 43, No. 43).
**75** Wickens, C. D. (2015). Noticing events in the visual workplace: The SEEV and NSEEV models. In: R. R. Hoffman, et al. eds. Part VI – Perception and Domains of Work and Professional Practice. Cambridge: Cambridge University Press, pp. 749-768.
**76** Wiese, S., Lotz, A., & Russwinkel, N. (2019). Seev-vm: Act-r visual module based on seev theory. In Proceedings of the 17th International Conference on Cognitive Modeling (pp. 301-307).


## 6.2 Research agenda to further the field

*Christian P. Janssen (Utrecht University, NL), Martin Baumann (Universität Ulm, DE), Shamsi Tamara Iqbal (Microsoft – Redmond, US), and Antti Oulasvirta (Aalto University, FI)*



During the conference all the attendees identified various areas that can be part of a longer-term research agenda to move the field forward. Below, we list some of these research agenda items very briefly. They are clustered within each of the broader challenges that were a core panel discussion. Of course, these challenges are only an incomplete subset of the many research questions that are out there.

## Challenge 1: What phenomena and driving scenarios need to be captured?

- Trust: For example, models could be used to study when is there (too much of) a deviation from an "appropriate level of trust"? But the deeper question is How to model / predict this in an online way?
- Mental models: How can models be used to aid people to have a good (more realistic, detailed) mental model of car's capabilities and feedback? And also vice-versa: how can the can have an appropriate mental model of the driver (and their knowledge).

## Challenge 2: What technical capabilities do computational models possess?

- In what areas can models be part of the driving system and actually be used on the road?
- What are the requirements and constraints from a legal perspective when models are used on the road?
- Different models have different goals. There has been discussion whether we need computational models to model actual driving (i.e., to be able to take control of a vehicle, or be a "digital twin" of the driver of a vehicle), or whether models are mostly used for understanding human behavior and during the design phases?
- A related question is then: Where can or should a driving model be a black box or white box model?

## Challenge 3: How can models benefit from advances in AI while avoiding pitfalls?

- All models have some mechanisms that link inputs to outputs. However, the components come from different places: theory ("white box"), data ("black box"), or a combination? Each technique has advantages and disadvantages. How can techniques and insights from white box and black box models be best combined?
- This ties in with a more general challenge of how to best balance between generalizability /variability and fixed structure of a model. What is truly fixed (and well represented in a model)? What is learned / variable? How can one know they have a properly generalizable model?
- Crafting a model that has both white box and black box items can also be seen as a scientific process or method in itself. This method can be more standardized.
- Researchers make quite some choices during the model selection (and developing) process. How can this be approached in a more principled manner?

## Challenge 4:What insights are needed for and from empirical research?

- For what problems are large datasets needed? And for what problems are small datasets sufficient? (i.e., balancing also more traditional statistical techniques with machine learning techniques)
- What is needed to then make correct inferences on both? Small datasets and big datasets each have their charms, but require different techniques and insights. Just because a dataset is larger, does not make its quality better nor does it mean that the inferences are more reliable (as some form of data quality needs to be ensured).
- Some form of standardized data set might be useful for model development and model competition. A research objective can be to develop and grow such a data set (and let it grow over time).
- The above is tied to a need to have benchmark tests / phenomena to test models on. Is there perhaps a "golden standard" test?

## Challenge 5: How can models inform design and governmental policy?

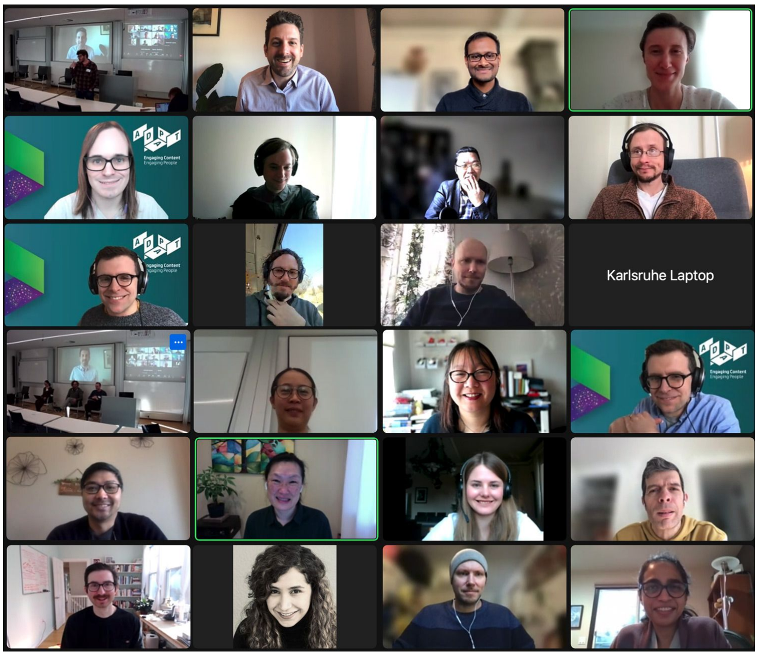
- How can cognitive models inform (the design of) future interaction best? Design efforts often explore specific scenarios, but within that look at various alternative designs. By comparisons, models sometimes are more fixed towards specific methods or outcomes. How can they also incorporate that flexibility?
- And how can the appropriate (modelling / model sketching) tools be made?

## Participants

- Martin Baumann
Universität Ulm, DE
- Jelmer Borst
University of Groningen, NL
- Alexandra Bremers
Cornell Tech – New York, US
- Duncan Brumby
University College London, GB
- Debargha Dey
TU Eindhoven, NL
- Patrick Ebel
Universität Köln, DE
- Martin Fränzle
Universität Oldenburg, DE
- Luisa Heinrich
Universität Ulm, DE
- Moritz Held
Universität Oldenburg, DE
- Jussi Jokinen
University of Jyväskylä, FI
- Dietrich Manstetten
Robert Bosch GmbH – Stuttgart, DE
- Gustav Markkula
University of Leeds, GB
- Roderick Murray-Smith
University of Glasgow, GB
- Antti Oulasvirta
Aalto University, FI
- Nele Rußwinkel
TU Berlin, DE
- Shadan Sadeghian Borojeni
Universität Siegen, DE
- Hatice Sahin
Universität Oldenburg, DE
- Philipp Wintersberger
TU Wien, AT
- Fei Yan
Universität Ulm, DE

## Remote Participants

- Linda Ng Boyle
University of Washington – Seattle, US
- Lewis Chuang
LMU München, DE
- Benjamin Cowan
University College – Dublin, IE
- Birsen Donmez
University of Toronto, CA
- Justin Edwards
ADAPT Centre – Dublin, IE
- Mark Eilers
Humatects – Oldenburg, DE
- Shamsi Tamara Iqbal
Microsoft – Redmond, US
- Christian P. Janssen
Utrecht University, NL
- Myounghoon Jeon
Virginia Polytechnic Institute – Blacksburg, US
- Xiaobei Jiang
Beijing Institute of Technology, CN
- Wendy Ju
Cornell Tech – New York, US
- Tuomo Kujala
University of Jyväskylä, FI
- Andrew Kun
University of New Hampshire – Durham, US
- Otto Lappi
University of Helsinki, FI
- Nikolas Martelaro
Carnegie Mellon University – Pittsburgh, US
- Andreas Riener
TH Ingolstadt, DE
- Boris van Waterschoot
Rijkswaterstaat – Utrecht, NL
- Yiqi Zhang
Pennsylvania State University – University Park, US


Karlsruhe Laptop

Report from Dagstuhl Seminar 22111

# Database Indexing and Query Processing


**Renata Borovica-Gajic*[1], Goetz Graefe*[2], Allison Lee*[3], Caetano Sauer*[4], and Pinar Tözün*[5]**

1 **The University of Melbourne, AU.** renata.borovica@unimelb.edu.au
2 **Google – Madison, US.** goetz.graefe@gmail.com
3 **Snowflake – San Mateo, US.** allison@snowflake.com
4 **Salesforce – München, DE.** caetano.sauer@salesforce.com
5 **IT University of Copenhagen, DK.** pito@itu.dk



**Abstract**

The Dagstuhl Seminar 22111 on "Database indexing and query processing", held from March 13 to March 18 2022, brought together researchers from academia and industry to discuss robustness in database management systems. This seminar was a continuation of previous seminars on the topic of Robust Query Processing, where we included *indexing* as a general topic and also discussed aspects that have not been addressed by the previous instances of the seminar. This article summarizes the main discussion topics, and presents the summary of the outputs of three work groups that discussed: i) storage architectures, ii) robust operators, and iii) indexing for data warehousing.



**Seminar** March 13–18, 2022 – http://www.dagstuhl.de/22111
**2012 ACM Subject Classification** Information systems → Data management systems; Information systems → Storage architectures
**Keywords and phrases** database, execution, hardware, optimization, performance, query
**Digital Object Identifier** 10.4230/DagRep.12.3.82


## 1 Executive Summary


*Renata Borovica-Gajic (The University of Melbourne, AU)*
*Goetz Graefe (Google – Madison, US)*
*Allison Lee (Snowflake – San Mateo, US)*
*Caetano Sauer (Salesforce – München, DE)*
*Pinar Tözün (IT University of Copenhagen, DK)*




The Dagstuhl Seminar 22111 on "Database indexing and query processing" assembled researchers from industry and academia for the fourth time to discuss robustness issues in database query performance. The seminar gathered researchers around the world working on indexing, storage, plan generation and plan execution in database query processing, and in cloud-based massively parallel systems with the purpose to address the open research challenges with respect to the robustness of database management systems. Delivering robust query performance is well known to be a difficult problem for database management systems. All experienced DBAs and database users are familiar with sudden disruptions in data centers due to poor performance of queries that have performed perfectly well in the past. The goal of the seminar was to discuss the current state-of-the-art, to identify specific research opportunities in order to improve the state-of-affairs in query processing, and to develop new

* Editor / Organizer




Database Indexing and Query Processing, *Dagstuhl Reports*, Vol. 12, Issue 3, pp. 82–96
Editors: Renata Borovica-Gajic, Goetz Graefe, Allison Lee, Caetano Sauer, and Pinar Tözün
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

approaches or even solutions for these opportunities, building upon successes of the past Dagstuhl Seminars [1, 2, 3]. The organizers (Renata Borovica-Gajic, Goetz Graefe, Allison Lee, Caetano Sauer, and Pinar Tözün) this time attempted to have a focused subset of topics that the participants discussed and analyzed in more depth. From the proposed topics on algorithm choices, join sequences, learned and lightweight indexes, database utilities, modern storage hardware, and benchmarking for robust query processing, the participants formed three work groups: i) one discussing indexing for data warehousing, ii) one discussing robust query operators, and iii) one discussing robust storage architectures. Upon choosing the topics of interest, the organizers then guided the participants to approach the topic through a set of steps: by first considering related work in the area; then introducing metrics and tests that will be used for testing the validity and robustness of the solution; after metrics, the focus was on proposing specific mechanisms for the proposed approaches; and finally the last step focused on the implementation policies. At the end of the week, each group presented their progress with the hope to continue their work towards a research publication. The reports of work groups are presented next.

## References


1. Renata Borovica-Gajic, Stratos Idreos, Anastasia Ailamaki, Marcin Zukowski, and Campbell Fraser. Smooth scan: Statistics-oblivious access paths. In Johannes Gehrke, Wolfgang Lehner, Kyuseok Shim, Sang Kyun Cha, and Guy M. Lohman, editors, *ICDE*, pages 315–326. IEEE Computer Society, 2015.
2. Renata Borovica-Gajic, Stratos Idreos, Anastasia Ailamaki, Marcin Zukowski, and Campbell Fraser. Smooth scan: robust access path selection without cardinality estimation. *VLDB J.*, 27(4):521–545, 2018.
3. Martin L. Kersten, Alfons Kemper, Volker Markl, Anisoara Nica, Meikel Poess, and Kai-Uwe Sattler. Tractor pulling on data warehouses. In Goetz Graefe and Kenneth Salem, editors, *DBTest*, page 7. ACM, 2011.

## 2 Table of Contents

**Executive Summary**
*Renata Borovica-Gajic, Goetz Graefe, Allison Lee, Caetano Sauer, and Pinar Tözün* 82

**Working Groups**

Storage Architectures
*Pinar Tözün, Goetz Graefe, Thomas Heinis, Sangjin Lee, Alberto Lerner, Danica Porobic, Daniel Ritter, Lukas Vogel, and Tianzheng Wang* . . . . . . . . . . . . . . . 85

Robust Operators
*Allison Lee, Carsten Binnig, Thomas Bodner, Campbell Fraser, Mhd Yamen Haddad, David Justen, Andrew Lamb, Bart Samwel, and Nga Tran* . . . . . . . . . . . . . . . 89

Indexing for Data Warehousing
*Caetano Sauer, Peter A. Boncz, Yannis Chronis, Jan Finis, Stefan Halfpap, Viktor Leis, Thomas Neumann, Anisoara Nica, Knut Stolze, and Marcin Zukowski* . . . . 92

**Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96

**Remote Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96

# 3 Working Groups

## 3.1 Storage Architectures

*Pinar Tözün (IT University of Copenhagen, DK), Goetz Graefe (Google – Madison, US), Thomas Heinis (Imperial College London, GB), Sangjin Lee (Hanyang University – Seoul, KR), Alberto Lerner (University of Fribourg, CH), Danica Porobic (Oracle Switzerland – Zürich, CH), Daniel Ritter (Hasso-Plattner-Institut, Universität Potsdam, DE), Lukas Vogel (TU München, DE), and Tianzheng Wang (Simon Fraser University – Burnaby, CA)*



The storage hierarchy has been getting deeper and more heterogeneous. In addition, platforms that enable computational storage and/or near-data processing are becoming more widely-available [1]. This storage landscape is an opportunity for data-intensive systems. However, it also presents us with several challenges when it comes to exploiting these technologies.

One key challenge that comes with the deeper and heterogeneous storage landscape is the various sources of unpredictability.

- Device types: Hard disks (HDD), Solid-state Drives (SSD), Persistent Memory (PMEM), DRAM have different device characteristics requiring the end-users to adopt different system optimizations. In addition, there may even be variety among the same class of devices. For example, SSDs are not a uniform class of devices. There are space-optimized (QLC, TLC) or speed-optimized (SLC) SSDs, and devices from different vendors behave differently.
- Interfaces: With the variety of the devices comes also the variety of device interfaces to interact with. Even within the same class of devices, there could be different options. For example, NVMe standard defines different interfaces for key-value SSDs, zoned-namespaces, computational storage (currently being standardized), etc.
- Disaggregated storage: It is common practice to separate compute nodes from storage nodes for large-scale hardware deployments. Accessing a locally-attached storage device could behave differently than accessing a storage device over the network.
- Access modes: There are different ways to access storage devices. One may include CPU on the path or bypass it using direct memory access (DMA). Some accesses may be transparent to the end-user implicitly being controlled by hardware itself, while some hardware may give more explicit controls to the end-user for application-specific optimizations.
- Workloads: Data-intensive workloads exhibit a high variety as well. While some workloads have well-behaved and predictable data read/write and movement characteristics, some can have unpredictable ad-hoc behavior.
- Infrastructure: Today many data-intensive systems run on the cloud. Cloud infrastructures take away the burden of managing a big hardware infrastructure from the end-users. However, they do so by abstracting or virtualizing hardware. This means that servers and storage devices used by a data-intensive system may change at any point. In addition, servers from different popular vendors that make up the cloud support different technologies. For example, Intel servers come with support for persistent memory, while AMD servers don't have this support. In contrast, AMD servers are optimized for supporting many PCIe lanes making them good choices if one wants to deploy many SSDs.

In the storage working group of this Dagstuhl Seminar, we specifically focused on the following research question: How can we robustly exploit the modern storage hierarchy despite all the sources of unpredictability?

If one digs deeper, at the heart of this challenge lies the challenge of orchestrating the data movement across the variety storage layers and devices. Therefore, the question above boils down to how can we orchestrate the data movement across layers to get more predictably good performance (a) when a workload is well-behaved and (b) when a workload isn't well-behaved?

### 3.1.1 Well-behaved scenario

We started our discussion focusing on the easier case, which is the well-behaved scenario. A representative workload for this scenario is external sort, which is a building block for many data-intensive operations such as the compaction operation for log-structured merge trees, sorting results of a query, etc. The key challenge with this operation is the temporary data it creates, which in turn creates storage pressure. Our goal is to design a robust and efficient external sort mechanism that can exploit different storage layers. The main design principle / intuition of our mechanism is to **separate the read and write traffic for the data movement**.

While we aimed at avoiding any assumptions regarding the functionality of available storage devices, one key requirement for the efficiency of our mechanism is having a form of DMA support. This is not an unreasonable requirement for today's storage landscape considering the availability of PCIe DMA engine for SSDs, ioat for moving data from DRAM to PMEM, S3 async put in the cloud, remote direct memory access (RDMA), etc.

Next we describe the external sort mechanism following our goal and design principle. There are two versions of it that differ in the way the sort and merge tasks are scheduled. Each version also has an associated illustration.

**Way up / Sorts:** The data to be sorted is read directly to processor caches from the persistent storage unit, which is the data source, using the DMA engine. The size unit of these fetches, let's call them runs, could be based on the LLC cache size per core divided by 2. The reason is for each direct data fetch to cache, even though one bypasses the CPU and memory layers, the associated memory space has to be allocated. We need twice the space to allow dual-buffering at LLC rather than going to DRAM while a core is sorting the fetched data.

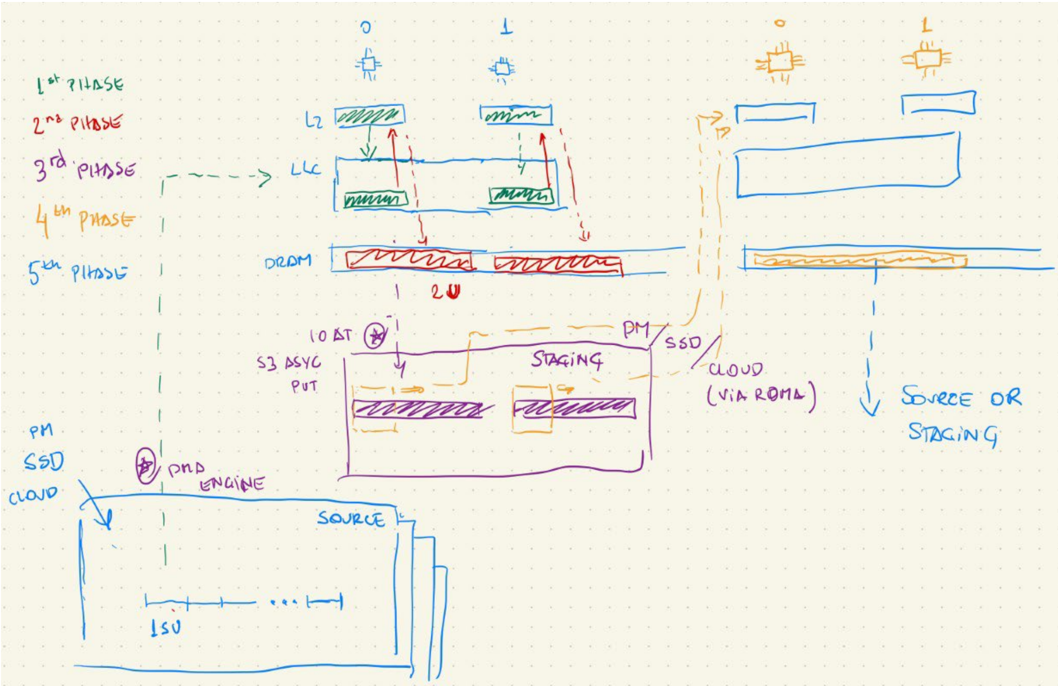
**Way down / Merges:** In the non-pipelined version, we first wait for all data to be sorted in units of runs before each core starts merging of these runs.

Each core performs merges till the DRAM size is exhausted. The merge-sorted run can be spilled to a persistent storage device as soon as the initial block/page of it is produced. This persistent storage device could be a different one than the data source if such a device is available. We will call it the staging area. Ideally, such a staging area should have low access latency such as PMEM or new-gen SSDs. Ideal number of runs a core merges at a time still requires a discussion.

The merge-sorted runs are read from the staging area using DMA using a fetch unit similar to the way up / sorting phase. However, this time, the runs are already sorted, so the cores just perform merging. This is repeated as long as it is needed to get the final merge-sorted run.

Where or which device we end up writing the sorted run to depends on the use case.

The main difference between the non-pipelined and the pipelined mechanisms is the way in which available cores are assigned to sort and merge tasks of the external merge-sort task.




■ **Figure 1** Not-pipelined scenario.

In the non-pipelined merge-sort mechanism, the sort (way-up) and merge (way-down) stages are separate stages. First, all the available cores sort the runs and then they all merge the sorted runs. Rather than this strict separation of the two stages, one can assign some cores for sorting the runs and some cores for merging, where the sorted runs are transferred to the cores responsible for merging. In this scheme the sort and merge operations are pipelined in stages, similar to earlier work like StagedDB and SharedDB.

How the data is fetched from or written to persistent storage devices is the same across the not-pipelined and pipelined versions of the merge-sort.

### 3.1.2 Not well-behaved scenario

During the seminar, we didn't have time to talk in detail about the not well-behaved scenarios. Such scenarios are characterized by the unpredictability of the read and write patterns such as online transaction processing (OLTP) workloads.

In the literature, a common way to handle OLTP workloads is creating hardware-conscious data structures such as log-structured merge trees [3], B-epsilon tree [2], Plush [4], Apex [5], etc. The main design goal when it comes to creating these data structures is to morph the workload's unpredictable data access patterns or movement to a more well-behaved pattern for the target storage device. The issue is that usually there is only one or two devices targeted such as DRAM & PMEM and DRAM & SSD. There are only a few recent works (e.g., Umzi [6], NovaLSM [7], etc.) that target multiple layers of storage hierarchy or disaggregated storage.

We overall support the approach of morphing the data movement patterns using novel and hardware-conscious data structures for not well-behaved workloads. On the other hand, we encourage our research community to consider the new and multiple layers of the storage hierarchy when adopting this approach.

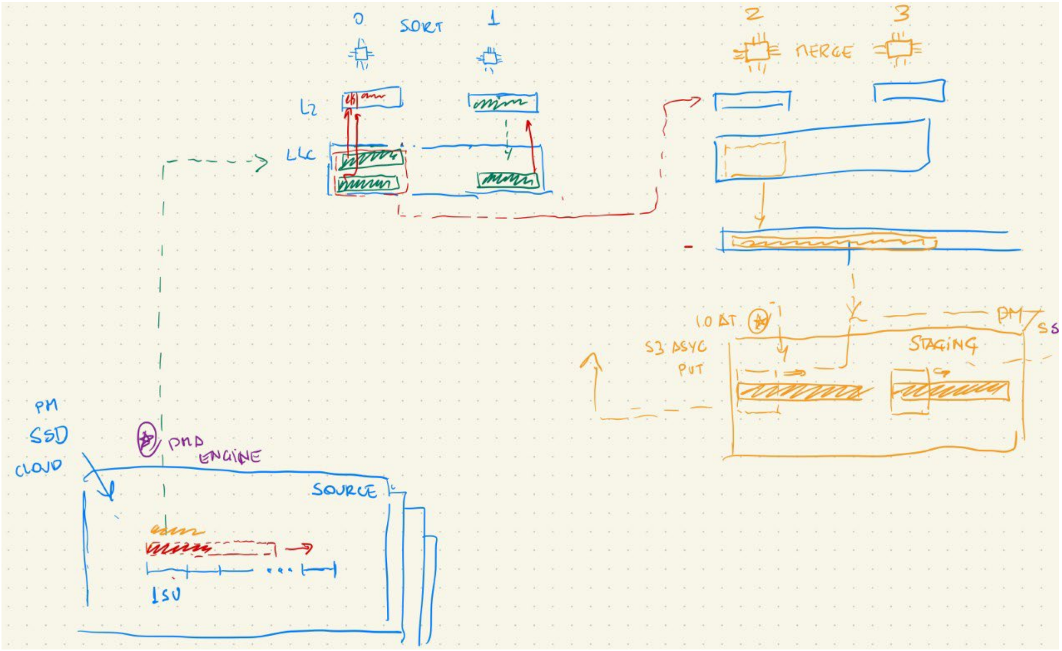



**Figure 2** Pipelined scenario.

### 3.1.3 Co-design of storage and data-intensive systems

During the seminar, we also didn't have the time to touch upon challenges for co-design and utilizing computational storage for data-intensive systems.

The co-design challenge boils down to the trade-off between having a predictable but sub-optimal mechanism vs unpredictable but smart mechanism. It is easier to have co-design principles for well-behaved workloads that would lead to predictably smart and optimal choices. However, the not well-behaved patterns may lead to unpredictability, which may overweigh the gains of being smart and optimal most of the time when interacting with storage devices. Nevertheless, it is still worthwhile to deploy storage and data-intensive system co-design mechanisms for operations such as filtering, encryption, compression, etc.

### 3.1.4 Next Steps

The next steps to this work are:

- Modeling the data movement cost to reason about benefits
- Reasoning about the tuning of parameters such as data fetch units, degree of parallelism, number of runs to merge, etc.
- Discussion on what happens if the server is shared with other requests
- Implementing the two versions of the external sort mechanism
- Additional work orthogonal to external sort design: extensive storage access tracing for big database systems

#### References

1 Alberto Lerner and Philippe Bonnet. Not your grandpa's SSD: the era of co-designed storage devices. In *SIGMOD*, pages 2852–2858, 2021.

2 Michael A. Bender, Martin Farach-Colton, William Jannen, Rob Johnson, Bradley C. Kuszmaul, Donald E. Porter, Jun Yuan, and Yang Zhan. An introduction to b$\epsilon$-trees and write-optimization. *login Usenix Mag.*, 40(5), 2015.
3 Patrick E. O'Neil, Edward Cheng, Dieter Gawlick, and Elizabeth J. O'Neil. The log-structured merge-tree (lsm-tree). *Acta Informatica*, 33(4):351–385, 1996.
4 Under submission/review.
5 Baotong Lu, Jialin Ding, Eric Lo, Umar Farooq Minhas, and Tianzheng Wang. APEX: A high-performance learned index on persistent memory. *Proc. VLDB Endow.*, 15(3):597–610, 2021.
6 Chen Luo, Pinar Tözün, Yuanyuan Tian, Ronald Barber, Vijayshankar Raman, and Richard Sidle. Umzi: Unified multi-zone indexing for large-scale HTAP. In *EDBT*, pages 1–12, 2019.
7 Haoyu Huang and Shahram Ghandeharizadeh. Nova-lsm: A distributed, component-based lsm-tree key-value store. In Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava, editors, *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 749–763. ACM, 2021.

## 3.2 Robust Operators

*Allison Lee (Snowflake – San Mateo, US), Carsten Binnig (TU Darmstadt, DE), Thomas Bodner (Hasso-Plattner-Institut, Universität Potsdam, DE), Campbell Fraser (Google – Mountain View, US), Mhd Yamen Haddad (Ecole Polytechnique – Palaiseau, FR), David Justen (Hasso-Plattner-Institut, Universität Potsdam, DE), Andrew Lamb (InfluxData – Boston, US), Bart Samwel (Databricks – Amsterdam, NL), and Nga Tran (InfluxData – Boston, US)*



The Dagstuhl Seminar 17222 on "Robust Performance in Database Query Processing" proposed a novel dynamic join order selection path method named "Plan of Least Resistance", which is described in the Dynamic Join Sequence working group section of the seminar report [1].

This novel algorithm aimed to increase the robustness of query processing by dynamically avoiding poorly chosen join orders based on runtime feedback. However it was not clear after the conclusion of Seminar 17222 how widely applicable and implementable this novel algorithm is.

Research Question: Is the "Plan of Least Resistance" (POLR) approach for robust query processing practical for commercial systems?

**Success definition**: Outline a minimal commercially viable implementation of POLR.

### 3.2.1 Review: Plan of Least Resistance

There are many open questions to this approach, so we focused only on those that must be resolved for a minimum viable commercial implementation of this approach:

- What shapes/orders of join plans are possible in the potentially routable paths, and which possible join orders should be included in the plan?
- What is the routing policy for the Multiplexer, and what cost metrics are required to implement that policy?

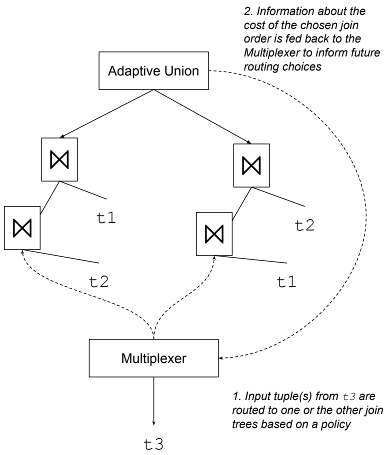



**Figure 3** Schematic Plan of Least Resistance: a Router chooses based on some model which of two join orders to route tuples from t3. One path joins t2 first and then t3, and one path joins t3 first followed by t2. The execution engine tracks the cost of evaluating the join tree that was chosen, for the tuple(s) and feeds that cost information back to the router to inform its future choices.

### 3.2.2 Join Order Selection

The goal of Join Order selection is to determine a practical way to pick join orders that will provide robust query performance for the switcher.

Assumptions:

- System that will use hash join with some form of sideways information passing (SIPS) filter that can be applied to join keys during scans of other relations
- Only consider linear join plans (left deep / right deep depending on which side you prefer to draw the build input)

Using the information about the input relation sizes, the system leverages a classical optimizer to pick a candidate set of plans. An algorithm was proposed to generate a set of plans that provides good coverage across the space of possible cardinalities of intermediate results. The initial table to scan can also be chosen with the help of the classical optimizer, to pick the best cost table that is robust across the space of cardinalities.

This approach is more robust than a fixed join order, and it is implementable in typical commercial database systems.

### 3.2.3 Multiplexer Tuple Routing

The goal of the router is to route input tuples the best among available join orders. Its dynamic nature also allows it to adapt automatically to changes in the input and be robust to various clusterings of input values in the incoming data stream.

We propose a “bounded regret” approach algorithm to select which possibility a particular tuple is routed to. Specifically, the user provides a budget for how much extra time the robust plan to spend vs. the fastest currently known plan. The router will then choose a tuple routing to stay within this budget.

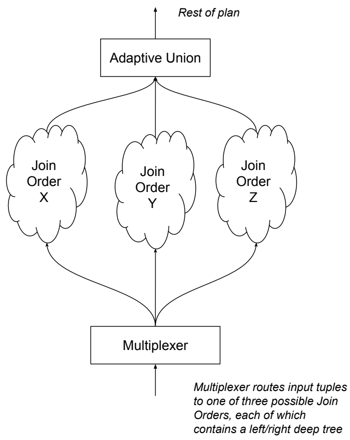



**Figure 4** A Multiplexer beneath an Adaptive Union routes tuples to one of three possible join orders.

Initially the router will send "enough" tuples to all three of the branches to be confident in the observed cost. After this initial phase, the router will send tuples to each join order with a certain probability, depending on the observed cost of that plan, in order to constrain the overhead to within the budget, while maximizing observations of potentially better plans. If the observed costs change significantly over time, then this algorithm is run again to update the weights.

This technique is more robust than picking the best order after initial measurement because it can switch between multiple plans over time, and even if it gets it wrong initially the runtime feedback loop can guide it to a better plan over time. This technique is implementable as it requires straightforward calculations that are neither overly burdensome to implement and require trivial CPU and memory resources, and are easy to test.

### 3.2.4 Next Steps

During our week at Dagstuhl, we proposed a practical, robust solution to join order selection in database systems. The next steps for this work include:

- Build a research prototype of our solution. This would allow us to experiment with some of the alternative policy options that we considered for join order generation and tuple routing.
- Propose solutions to open questions unrelated to the minimal implementation, including different join shapes, distributed execution plans, and spilling operators.

#### References


1 Renata Borovica-Gajic, Goetz Graefe, and Allison W. Lee. Robust performance in database query processing (dagstuhl seminar 17222). *Dagstuhl Reports*, 7(5):169–180, 2017.

## 3.3 Indexing for Data Warehousing

*Caetano Sauer (Salesforce – München, DE), Peter A. Boncz (CWI – Amsterdam, NL), Yannis Chronis (University of Wisconsin-Madison, US), Jan Finis (Salesforce – München, DE), Stefan Halfpap (Hasso-Plattner-Institut, Universität Potsdam, DE), Viktor Leis (Universität Erlangen-Nürnberg, DE), Thomas Neumann (TU München, DE), Anisoara Nica (SAP SE – Waterloo, CA), Knut Stolze (IBM Deutschland – Böblingen, DE), and Marcin Zukowski (Snowflake – San Mateo, US)*



Selective queries are quite common in large-scale data analytics; for example, when drilling down into a specific customer in a dashboard. Traditionally, selective queries are optimized by creating secondary indexes. However, because of their large size, expensive maintenance, and difficulty to tune and automate, indexes are typically not used in modern cloud data warehouses. Instead, such systems rely mostly on full table scans and lightweight optimizations like min/max filtering, whose effectiveness depends heavily on the data layout and value distributions. It is also difficult to predict whether certain columns will be targeted by selective queries or not, which may preclude an upfront decision to create indexes.

In this working group, we sketched a general indexing framework called SPA (Smooth Predicate Acceleration). It optimizes selective queries automatically, by adaptively indexing subsets of data in an incremental and workload-driven manner. It makes fine-granular decisions and continuously monitors their benefit, dynamically allocating an optimization budget in a way that bounds the additional cost of indexing. Furthermore, it guarantees a performance improvement in the cases where indexes—potentially partial ones—prove to be beneficial. On the other hand, when indexes lose their benefit due to a shifting workload, they are also gradually deconstructed in favor of optimizations that accommodate recent trends.

### 3.3.1 Desiderata

The framework envisioned in our working group should be:

- Workload-driven: indexes are created and dropped solely based on workload observations, without upfront decisions or manual interventions.
- Smooth: index maintenance is carried out in incremental steps, as a side-effect of table scans and without spikes in query latency.
- Economical: decisions are taken and evaluated based solely on the monetary cost in comparison to a baseline of full table scans.
- Cost-bounded (i.e., “do no harm”): bad decisions should not impact the user-observed response times by more than a configurable percentage.
- Modular: the framework supports different types of index or summary structures, and their individual characteristics are taken into account by the economic model.

### 3.3.2 General approach

The SPA framework observes the workload and automatically maintains partial indexes in an incremental manner. The decisions taken by the framework are purely economical: it tracks the additional cost of index maintenance (for both computing and storage) as well as the benefit provided by indexes during scans. A positive balance on this benefit gives the framework more budget to continue building indexes; a negative balance, on the other hand, leads to a gradual deconstruction of indexes. Thus, index creation can be seen as an investment with continuously evaluated returns. The additional cost of indexing is bounded thanks to a configurable budget (or "indexing tax"), which is specified as a percentage of the cost of a full table scan (e.g., 1%): if none of the indexing investments pay off, the system guarantees that queries will not be slowed down by more than this percentage on average.

Indexes are built incrementally by first indexing individual subsets of a table, such as a file or a block on storage. These are considered units of scanning which can be skipped with available summaries such as min/max small materialized aggregates (SMAs) [2, 1]. If the available summaries are not able to filter out a particular block and matches are not found for a given predicate, then the SPA framework might create an index on that block specifically. On subsequent scans, that index can be probed before fetching and scanning the corresponding block. As more and more blocks get indexed, they might also be merged into larger indexes covering multiple blocks, similar to a log-structured merge tree. These partial index structures might also lose their value over time, in which case a caching policy can drop them or deconstruct large indexes into smaller ones.

### 3.3.3 Simulation

To simulate the behavior of SPA, we implemented a mock of an in-memory, column-oriented table scan operator in C++. This prototype, available in an open source repository[1], organizes records into blocks of 100,00 tuples. It uses a simple randomized approach to create indexes in these blocks individually. This works as follows: whenever a block is scanned and no match is found for the simulated predicate, the system randomly chooses whether to create an index for this block. The probability of this choice is proportional to an artificial budget variable. This variable is incremented by a small fixed amount with every block scanned (1), and decremented by a much larger amount if an index is created (2). In the case where an index is available and this index allows a block to be skipped, the budget is increased by a comparatively large amount (3). The reasoning behind each of these budget changes is explained below, referring back to the numbers in parentheses above:

1. A small budget should be accrued regularly to allow for index creations in the first place; this can be seen as a regular small deposit (or savings) into the index maintenance account.
2. Creating an index has a non-negligible cost on scan performance; it is an investment that decreases the account balance but hopefully brings returns in the future.
3. If an index allows the scan operator to skip blocks, then the investment has paid off, and returns are deposited into the account.

As more budget is accrued (hopefully exponentially thanks to compounded returns) and more indexes are built, smaller indexes are also merged into larger ones. Just like index creation, the merge operation also deducts from the budget and pays back returns whenever it is used to avoid scan work.

Figure 5 below shows the observed query response times from an execution of this prototype with different deposit rates, i.e., the budget increase with every block scan in step 1 above. Note that this is a unitless parameter, as it just serves to scale the probability of creating an index. This experiment sends repeated queries (x axis) with a random equality predicate on a given column. The query response time is plotted in the y axis.

[1] `https://github.com/JFinis/dagstuhl-spa`

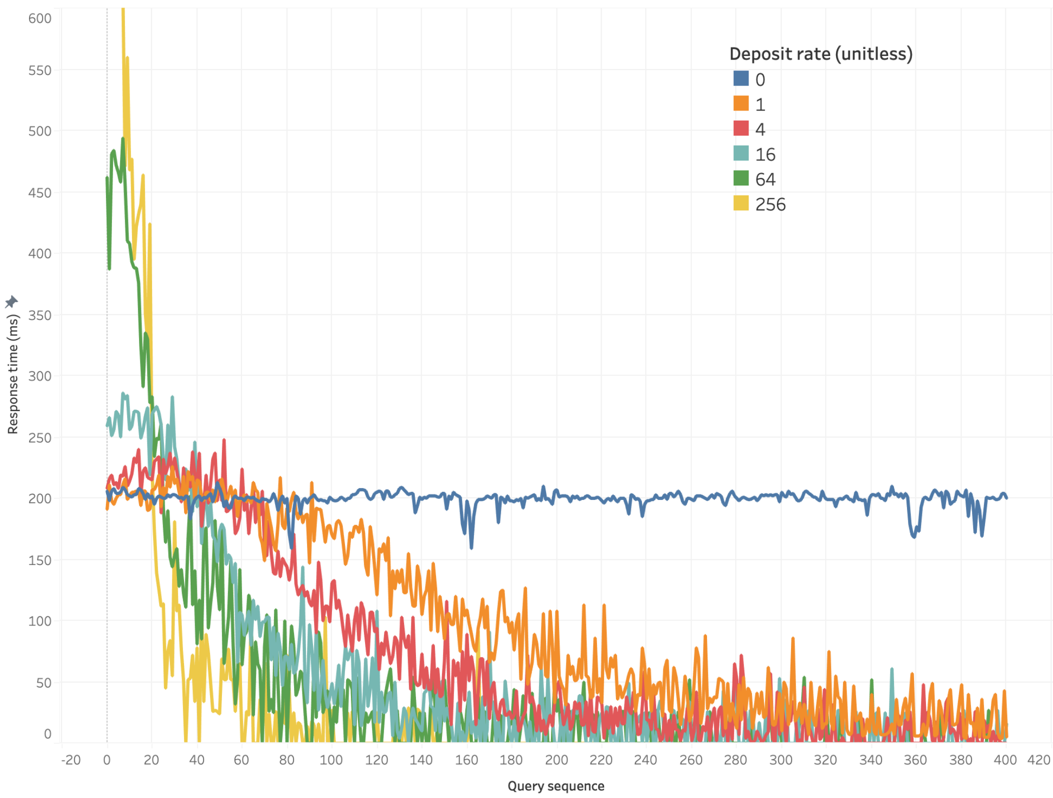



**Figure 5** Simulation of budget-driven index creation.

As the results show, a deposit rate of zero (blue line) has nearly constant response time of 200ms, serving as the baseline for the experiment. As the deposit rate increases, the first queries in the sequence observe larger response times, but they converge faster into a fully index-based scan, with response time under 50ms. This reflects the expected behavior of our economic model: lower deposit rates have lower disturbance in query response times, while higher deposit rates benefit faster from indexing performance; in the end, all choices converge to faster execution speeds thanks to indexing.

### 3.3.4 Future work

Our working group considers the ideas developed during this seminar novel and industry-relevant. As such, we plan to refine these ideas into a more detailed description of the SPA framework and submit them as part of a vision paper to a major database conference. To validate the benefits investigated with the prototype implementation described above, we also plan to implement a cost-based prototype in a commercial database system and publish our evaluation results as part of the aforementioned vision paper.

On the technical side, there are also multiple avenues to pursue in terms of design choices:

- Experiment with different index structures, which might trade-off accuracy for space consumption.
- Evaluate partial index structures in terms of how efficient and simple they are to merge and deconstruct incrementally (i.e., their composability).
- Investigate different cost models, especially focused on the cost of resources in the cloud.

## References


1 Goetz Graefe. Fast loads and fast queries. In Torben Bach Pedersen, Mukesh K. Mohania, and A Min Tjoa, editors, *DaWaK*, volume 5691 of *Lecture Notes in Computer Science*, pages 111–124. Springer, 2009.

2 Guido Moerkotte. Small materialized aggregates: A light weight index structure for data warehousing. In *VLDB*, pages 476–487. Morgan Kaufmann, 1998.

## Participants


- Carsten Binnig
TU Darmstadt, DE
- Thomas Bodner
Hasso-Plattner-Institut, Universität Potsdam, DE
- Peter A. Boncz
CWI – Amsterdam, NL
- Yannis Chronis
University of Wisconsin-Madison, US
- Jan Finis
Salesforce – München, DE
- Campbell Fraser
Google – Mountain View, US
- Mhd Yamen Haddad
Ecole Polytechnique – Palaiseau, FR
- Stefan Halfpap
Hasso-Plattner-Institut, Universität Potsdam, DE
- Thomas Heinis
Imperial College London, GB
- David Justen
Hasso-Plattner-Institut, Universität Potsdam, DE
- Andrew Lamb
InfluxData – Boston, US
- Allison Lee
Snowflake – San Mateo, US
- Sangjin Lee
Hanyang University – Seoul, KR
- Viktor Leis
Universität Erlangen-Nürnberg, DE
- Alberto Lerner
University of Fribourg, CH
- Thomas Neumann
TU München, DE
- Anisoara Nica
SAP SE – Waterloo, CA
- Danica Porobic
Oracle Switzerland – Zürich, CH
- Daniel Ritter
Hasso-Plattner-Institut, Universität Potsdam, DE
- Bart Samwel
Databricks – Amsterdam, NL
- Caetano Sauer
Salesforce – München, DE
- Knut Stolze
IBM Deutschland – Böblingen, DE
- Pinar Tözün
IT University of Copenhagen, DK
- Nga Tran
InfluxData – Boston, US
- Lukas Vogel
TU München, DE
- Tianzheng Wang
Simon Fraser University – Burnaby, CA
- Marcin Zukowski
Snowflake – San Mateo, US


## Remote Participants


- Renata Borovica-Gajic
The University of Melbourne, AU
- Goetz Graefe
Google – Madison, US

Report from Dagstuhl Seminar 22121

# 3D Morphable Models and Beyond


**James Gardner*[1], Bernhard Egger†[2], William Smith†[3], Christian Theobalt†[4], and Stefanie Wuhrer†[5]**

1 **University of York, GB.** james.gardner@york.ac.uk
2 **Friedrich-Alexander-Universität Erlangen-Nürnberg, DE.** bernhard.egger@fau.de
3 **University of York, GB.** william.smith@york.ac.uk
4 **MPI für Informatik – Saarbrücken, DE.** theobalt@mpi-inf.mpg.de
5 **INRIA – Grenoble, FR.** stefanie.wuhrer@inria.fr



**Abstract**

3D Morphable Models are models separating shape from appearance variation. Typically, they are used as a statistical prior in computer graphics and vision. Recent success with neural representations have caused a resurgence of interest in visual computing problems, leading to more accurate, higher fidelity, more expressive, and memory-efficient solutions. This report documents the program and the outcomes of Dagstuhl Seminar 22121, "3D Morphable Models and Beyond". This meeting of 39 researchers covered various topics, including 3D morphable models, implicit neural representations, physics-inspired approaches, and more. We summarise the discussions, presentations and results of this workshop.



**Seminar** March 20–25, 2022 – http://www.dagstuhl.de/22121

**2012 ACM Subject Classification** Computing methodologies → Computer graphics; Computing methodologies → Image-based rendering; Computing methodologies → Shape modeling; Computing methodologies → Animation; Computing methodologies → Computer vision; Computing methodologies → 3D imaging

**Keywords and phrases** 3D Computer Vision, Generative Models, Neural Rendering, Implicit Representations, Computer Graphics, Statistical Modelling

**Digital Object Identifier** 10.4230/DagRep.12.3.97


## 1 Executive Summary


*James Gardner (University of York, GB)*
*Bernhard Egger (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)*
*William Smith (University of York, GB)*
*Christian Theobalt (MPI für Informatik – Saarbrücken, DE)*
*Stefanie Wuhrer (INRIA – Grenoble, FR)*




A total of 63 people were invited to the seminar in the first round of invitations. 39 people attended, with 15 of those attending the seminar virtually. Participants came from both academia and industry and at varying stages of their careers. As this seminar took place at the trailing end of the Covid-19 pandemic, it ran in a hybrid format, and for many attendees, Dagstuhl was the first in-person seminar in several years. Due to the fantastic facilities of the Dagstuhl campus, the hybrid format was a great success, enabling accessible and inclusive

* Editorial Assistant / Collector
† Editor / Organizer



3D Morphable Models and Beyond, *Dagstuhl Reports*, Vol. 12, Issue 3, pp. 97–116
Editors: James Gardner, Bernhard Egger, William Smith, Christian Theobalt, and Stefanie Wuhrer
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

communication with remote participants. Daily Covid testing for those in-person ensured that everyone remained safe throughout the week. Eighteen presented their work in around 15-30 minute presentations; an abstract for each talk is included in this report.

Alongside traditional presentations, many sessions were left available for activities suggested by the seminar participants. These could involve workshops, discussions, presentations, or any other suggested format. During the week, participants could propose plans for the flexible sessions and the structure of the seminar became fixed as activities and topics for the sessions were provided. Summaries for the results of these flexible sessions are contained in this report. One slot was reserved for a joint group discussion on the ethical concerns of the research we are developing. This resulted in a vivid discussion on the steps we as a community should be taking to encourage the ethical use of the technology we are developing. One of the discussions that received broad support was the design of a cheap, open-source method for collecting camera calibrated illumination environments. This resulted in a Slack channel for the group of interested researchers and the pursuit of an early prototype design. We started the seminar with a short introduction from all participants. Everyone was given one slide to introduce themselves and asked to prepare a question, challenge or goal to discuss during the seminar.

## 2 Table of Contents

**Executive Summary**
*James Gardner, Bernhard Egger, William Smith, Christian Theobalt, Stefanie Wuhrer* 97

**Overview of Talks**

Learning to Fit Morphable Models
*Federica Bogo* . . . 101

From Pixels to Expressive 3D Bodies
*Timo Bolkart* . . . 101

Plausible (Neural) Rendering of Bodies & Garments in Motion
*Duygu Ceylan* . . . 102

Inferring people's anatomic skeleton from their external appearance
*Marilyn Keller* . . . 102

Deep Signatures – Learning Invariants of Planar Curves
*Ron Kimmel and Roy Velich* . . . 102

Computer Vision does not generalize – 3DMMs and beyond can help
*Adam Kortylewski* . . . 103

NeRF for View Synthesis
*Ben Mildenhall* . . . 104

Deep Relighting of 3D Faces
*Shunsuke Saito* . . . 104

Digital Humans in Motion
*Justus Thies* . . . 105

A Structured Latent Space for Human Body Motion Generation
*Stefanie Wuhrer* . . . 105

Implicit 3DMMs for Full Heads Including Hair
*Tarun Yenamandra* . . . 106

Towards Precise Completion of Deformable Shapes
*Oshri Halimi* . . . 106

Do We Still Need to Detect Faces and Facial Landmarks? Do We Need to Estimate 3D Face Shapes?
*Tal Hassner* . . . 107

On Implicit Avatars, Racial Bias, and the Light/Albedo Ambiguity
*Victoria Fernández Abrevaya* . . . 107

**Working groups**

What are the 'Killer Applications' of 3D Implicit Representations
*Ben Mildenhall* . . . 108

Are Neural Implicit Representations the Future of Morphable Models?
*Christian Theobalt* . . . 109

How Much 3D is Needed?
*Duygu Ceylan* . . . 109

Metrical 3D Reconstruction of the Human Face
*Justus Thies* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

Synthetic Data Generation Using Morphable Models
*Federica Bogo* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110

Morphable Models from Physics
*Dan Casas* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111

Ethics
*Bernhard Egger, William Smith, Christian Theobalt, Stefanie Wuhrer* . . . . . . . . 112

**Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115

**Remote Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116

# 3 Overview of Talks

## 3.1 Learning to Fit Morphable Models

*Federica Bogo (Meta Reality Labs Research – Zürich, CH)*


**Joint work of** Federica Bogo, Vasileios Choutas, Jingjing Shen, Julien Valentin

Fitting parametric models of human bodies, hands or faces to image data is an important problem in computer vision. Many recent approaches leverage deep neural networks to regress the parameters of the model directly from the input. These methods are fast and robust, but require large amounts of annotated data and may fail to tightly fit the observations. Therefore, their output is often leveraged as a starting point for an iterative, optimization-based algorithm minimizing an energy function. These functions typically involve a data term, plus priors encoding knowledge of the problem's structure; unfortunately they are difficult to both formulate and tune. In this talk, I will discuss how learning-based continuous optimization can capture the best of both deep-learning-based regression and classic optimization. I will discuss recent advances in the field and introduce a novel, learning-based approach for human body fitting, inspired by the Levenberg-Marquardt algorithm. Finally, I will identify some limitations of current state-of-the-art approaches and outline a few directions for future research.

## 3.2 From Pixels to Expressive 3D Bodies

*Timo Bolkart (MPI für Intelligente Systeme – Tübingen, DE)*


**Joint work of** Radek Danecek, Michael J. Black, Timo Bolkart, Yao Feng, Vasileios Choutas, Dimitrios Tzionas
**Main reference** Radek Danecek, Michael J. Black, Timo Bolkart: "EMOCA: Emotion Driven Monocular Face Capture and Animation", CoRR, Vol. abs/2204.11312, 2022.
**URL** https://doi.org/10.48550/arXiv.2204.11312
**Main reference** Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, Michael J. Black: "Collaborative Regression of Expressive Bodies using Moderation", in Proc. of the International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021, pp. 792–804, IEEE, 2021.
**URL** https://doi.org/10.1109/3DV53792.2021.00088

Recovering expressive humans from images is essential for understanding human behavior. Faces and their emotional expressions provide an important source of information about a person's internal emotional state. Unfortunately, the best recent 3D face regression methods from monocular images are unable to capture the full spectrum of facial expression, such as subtle or extreme emotions. We address this problem with EMOCA, by introducing a novel deep perceptual emotion consistency loss during training, which helps ensure that the reconstructed 3D expression matches the expression depicted in the input image. Reasoning about humans in images requires estimating not only the face, but the full expressive body. To that end, we present PIXIE. PIXIE combines a body-driven attention scheme with a moderator that merges features of body-part experts to reconstruct 3D bodies with articulated hands and expressive faces directly from images.

## 3.3 Plausible (Neural) Rendering of Bodies & Garments in Motion

*Duygu Ceylan (Adobe Research – London, GB)*


**Joint work of** Duygu Ceylan, Meng Zhang, Niloy J. Mitra, Tuanfeng Wang, Jae Shin Yoon, Cynthia Lu, Jimei Yang, Hyun Soo Park, Zhixin Shu

While there has been a lot of work on capturing 3D human body pose from single images and learning to generate pose-conditioned human models either in 2D or 3D, a relatively less explored area is to model motion dependent deformations. In this talk, I will discuss some of my recent work in utilizing motion features to synthesize plausible garments in 2D or 3D. I will specifically point out the main challenges and speculate on potential directions.

## 3.4 Inferring people's anatomic skeleton from their external appearance

*Marilyn Keller (MPI für Intelligente Systeme – Tübingen, DE)*


**Joint work of** Marilyn Keller, Silvia Zuffi, Michael J. Black, Sergi Pujades
**Main reference** Marilyn Keller, Silvia Zuffi, Michael J. Black, Sergi Pujades: "OSSO: Obtaining Skeletal Shape from Outside", CoRR, Vol. abs/2204.10129, 2022.
**URL** https://doi.org/10.48550/arXiv.2204.10129

Modeling the human internal anatomy is key in medicine and biomechanics. While many statistical models of the human being have been developed, those mainly describe their external appearance or individual bones. In this talk, I present how we learn a statistical model of the whole skeleton and its correlation with the body shape.

We do so using 1000 male and 1000 female dual-energy X-ray absorptiometry (DXA) scans. To these, we fit a parametric 3D body shape model (STAR) to capture the body surface and a novel part-based 3D skeleton model to capture the bones. This provides inside/outside training pairs. We model the statistical variation of full skeletons using PCA in a pose-normalized space. We then train a regressor from body shape parameters to skeleton shape parameters and refine the skeleton to satisfy constraints on physical plausibility. We name our inference tool OSSO, for "Obtaining Skeletal Shape from Outside". Given an arbitrary 3D body shape and pose, OSSO predicts a realistic skeleton inside.

## 3.5 Deep Signatures – Learning Invariants of Planar Curves

*Ron Kimmel (Technion – Haifa, IL) and Roy Velich (Technion – Haifa, IL)*



According to an important theorem by É. Cartan [1, 2], two planar curves are related by a group action, if and only if their signature curves, with respect to a given transformation group, are identical. Signature curves are parametrized by the group's differential invariants. Therefore, differential invariants provide a fundamental building block for the solution of the equivalence problem of planar curves, and geometric structures in general. We propose a learning paradigm for numerical approximation of differential invariants of planar curves.

Deep neural-networks' (DNNs) universal approximation properties are utilized to estimate geometric measures. The proposed framework is shown to be a preferable alternative to axiomatic constructions. Specifically, we show that DNNs can learn to overcome instabilities and sampling artifacts and produce numerically-stable signatures for curves subject to a given group of transformations in the plane. We compare the proposed schemes to alternative state-of-the-art axiomatic constructions of group invariant arc-lengths and curvatures. We evaluate our models qualitatively and quantitatively and propose a benchmark dataset to evaluate approximation models of differential invariants of planar curves.

#### References

1 E. Cartan. *La méthode du repère mobile, la théorie des groupes continus et les espaces généralisés*. The Mathematical Gazette, 1935

2 Olver, Peter J. *Classical Invariant Theory*. Cambridge University Press, 1999

## 3.6 Computer Vision does not generalize – 3DMMs and beyond can help

*Adam Kortylewski (MPI für Informatik – Saarbrücken, DE)*


**Main reference** Angtian Wang, Adam Kortylewski, Alan L. Yuille: "NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation", CoRR, Vol. abs/2101.12378, 2021.
**URL** https://arxiv.org/abs/2101.12378

In this talk, I pointed out a fundamental issue in computer vision research, namely that there is a large gap between the performance of vision models on academic benchmarks and their generalization ability in real-world applications. As an illustrative example, I contrasted the outstanding performance of vision models on popular and challenging benchmarks with the still unsolved problem of detecting simple STOP signs with self-driving cars. The fundamental issue is that we assume in academic benchmarks that the training and test data are very similar (i.e. i.i.d. distributed), while autonomous systems that interact with the real-world are often confronted with data that is, in some aspect, different from what has been observed at training time (e.g. unseen illumination, context, occlusion, texture, etc.). In the second half of this talk, I discussed how advances in statistical generative models and neural rendering could potentially help to close the genralization gap. I referred to some of our recent work on integrating deep neural networks with statistical generative models in 2D [2] and 3D-aware architectures [1], which enabled machines to generalize to unseen occlusion, to perform amodal segmentation [5], and to reason about occlusion ordering [3]. I also discussed recent work where we used generative models to benchmark vision systems through adversarial examination [4], and efforts to design new datasets that focus on capturing real-world generalization [6].

#### References

1 Wang, Angtian, Adam Kortylewski, and Alan Yuille. "NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation." International Conference on Learning Representations. 2021.

2 Kortylewski, Adam, et al. "Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

3 Yuan, Xiaoding, et al. "Robust instance segmentation through reasoning about multi-object occlusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
4 Ruiz, N., Kortylewski, A., Qiu, W., Xie, C., Bargal, S. A., Yuille, A., and Sclaroff, S. (2021). Simulated Adversarial Testing of Face Recognition Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
5 Sun, Yihong, Adam Kortylewski, and Alan Yuille. "Amodal Segmentation through Out-of-Task and Out-of-Distribution Generalization with a Bayesian Model". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
6 Zhao, Bingchen, et al. "ROBIN: A Benchmark for Robustness to Individual Nuisancesin Real-World Out-of-Distribution Shifts." arXiv preprint arXiv:2111.14341 (2021).

## 3.7 NeRF for View Synthesis

*Ben Mildenhall (Google – London, GB)*


**Joint work of** Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, Jonathan Barron, Peter Hedman, Dor Verbin, Todd Zickler
**Main reference** Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng: "NeRF: representing scenes as neural radiance fields for view synthesis", Commun. ACM, Vol. 65(1), pp. 99–106, 2022.
**URL** https://doi.org/10.1145/3503250

Recent years have seen a massive jump in quality for the task of photorealistic novel view synthesis, mainly driven by hybrid neural rendering pipelines in which part or all of the scene representation or rendering process are optimized for final image quality using gradient descent. Our group at Google has focused on pushing further towards higher resolution, bigger scenes, and more physically accurate view synthesis, with the hopes that progress on representations and rendering methods for this underpinning task can be fruitfully transferred to many other problems in 3D vision. I will give a brief overview of our recent work on extending NeRF to perform better on large scenes, shiny objects, and noisy camera data.

## 3.8 Deep Relighting of 3D Faces

*Shunsuke Saito (Reality Labs – Pittsburgh, US)*


**Joint work of** Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, Jason M. Saragih
**Main reference** Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, Jason M. Saragih: "Deep relightable appearance models for animatable faces", ACM Trans. Graph., Vol. 40(4), pp. 89:1–89:15, 2021.
**URL** https://doi.org/10.1145/3450626.3459829

We present a method for building high-fidelity animatable 3D face models that can be posed and rendered with novel lighting environments in real-time. Our main insight is that relightable models trained to produce an image lit from a single light direction can generalize to natural illumination conditions but are computationally expensive to render. On the other hand, efficient, high-fidelity face models trained with point-light data do not generalize to novel lighting conditions. We leverage the strengths of each of these two approaches. We first train an expensive but generalizable model on point-light illuminations, and use it to generate a training set of high-quality synthetic face images under natural illumination conditions. We then train an efficient model on this augmented dataset, reducing the generalization ability requirements. As the efficacy of this approach hinges on the quality of the synthetic data we can generate, we present a study of lighting pattern combinations for dynamic captures and evaluate their suitability for learning generalizable relightable models. Towards achieving the best possible quality, we present a novel approach for generating dynamic relightable faces that exceeds state-of-the-art performance. Our method is capable of capturing subtle lighting effects and can even generate compelling near-field relighting despite being trained exclusively with far-field lighting data.

## 3.9 Digital Humans in Motion

*Justus Thies (MPI für Intelligente Systeme – Tübingen, DE)*



The main theme of my work is to capture and to (re-)synthesize the real world using commodity hardware. It includes the modeling of the human body, tracking, as well as the reconstruction and interaction with the environment. The digitization is needed for various applications in AR/VR as well as in movie (post-)production. Teleconferencing and remote collaborative working in VR is of high interest since it is the next evolution step of how people communicate. A realistic reproduction of appearances and motions is key for such applications. Capturing natural motions and expressions as well as the photorealistic reproduction of images under novel views are challenging. With the rise of deep learning methods and, especially, neural rendering, we see immense progress to succeed in these challenges. In this talk, I will focus on the image synthesis of humans, the underlying representation of appearance, geometry, and motion to allow for explicit and implicit control over the synthesis process.

## 3.10 A Structured Latent Space for Human Body Motion Generation

*Stefanie Wuhrer (INRIA – Grenoble, FR)*


**Joint work of** Mathieu Marsot, Stefanie Wuhrer, Jean-Sebastien Franco, Stephane Durocher
**Main reference** Mathieu Marsot, Stefanie Wuhrer, Jean-Sebastien Franco, Stephane Durocher: "A structured latent space for human body motion generation", arXiv, 2021.
**URL** https://doi.org/10.48550/ARXIV.2106.04387

We study learning a structured latent space to represent and generate temporally and spatially dense 4D human body motion. Once trained, the proposed model generates a multi-frame sequence of dense 3D meshes based on a single point in a low-dimensional latent space. This latent motion representation can be learned in a data-driven framework that builds upon two existing lines of works. The first analyzes temporally dense skeletal data to capture the global displacement, poses and temporal evolution of the motion, while the second analyzes static densely captured human scans in 3D to represent realistic 3D human body surfaces in a low-dimensional space. Building upon the respective advantages of these two concepts allows the model to simultaneously represent temporal motion information for sequences of varying duration and detailed 3D geometry at every time instant of the motion. Experiments demonstrate that the resulting latent space is structured in the sense that similar motions form clusters in this space, and that the latent space allows to generate plausible interpolations between different actions.

## 3.11 Implicit 3DMMs for Full Heads Including Hair

*Tarun Yenamandra (TU München, DE)*


**Joint work of** Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, Christian Theobalt
**Main reference** Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, Christian Theobalt: "i3DMM: Deep Implicit 3D Morphable Model of Human Heads", in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pp. 12803–12813, Computer Vision Foundation / IEEE, 2021.
**URL** https://openaccess.thecvf.com/content/CVPR2021/html/Yenamandra_i3DMM_Deep_Implicit_3D_Morphable_Model_of_Human_Heads_CVPR_2021_paper.html

3DMMs are morphable models of human faces. Existing mesh-based 3DMMs consider only a part of the human head, commonly the face region. While some also model the shape of the head, no existing 3DMM can model hair along with other features of human heads. This is partly due to the unavailability of full head data and due to the challenges in modeling hair with mesh-based representations. Can an implicit representation-based 3DMM help solve some of the limitations? What are the challenges of such models?

## 3.12 Towards Precise Completion of Deformable Shapes

*Oshri Halimi (Technion – Haifa, IL)*


**Joint work of** Oshri Halimi, Ido Imanuel, Or Litany, Giovanni Trappolini, Emanuele Rodolà, Leonidas J. Guibas, Ron Kimmel
**Main reference** Oshri Halimi, Ido Imanuel, Or Litany, Giovanni Trappolini, Emanuele Rodolà, Leonidas J. Guibas, Ron Kimmel: "Towards Precise Completion of Deformable Shapes", in Proc. of the Computer Vision – ECCV 2020 – 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV, Lecture Notes in Computer Science, Vol. 12369, pp. 359–377, Springer, 2020.
**URL** https://doi.org/10.1007/978-3-030-58586-0_22

According to Aristotle, "the whole is greater than the sum of its parts". This statement was adopted to explain human perception by the Gestalt psychology school of thought in the twentieth century. Here, we claim that when observing a part of an object which was previously acquired as a whole, one could deal with both partial correspondence and shape completion in a holistic manner. More specifically, given the geometry of a full, articulated object in a given pose, as well as a partial scan of the same object in a different pose, we address the new problem of matching the part to the whole while simultaneously reconstructing the new pose from its partial observation. Our approach is data-driven and takes the form of a Siamese autoencoder without the requirement of a consistent vertex labeling at inference time; as such, it can be used on unorganized point clouds as well as on triangle meshes. We demonstrate the practical effectiveness of our model in the applications of single-view deformable shape completion and dense shape correspondence, both on synthetic and real-world geometric data, where we outperform prior work by a large margin.

## 3.13 Do We Still Need to Detect Faces and Facial Landmarks? Do We Need to Estimate 3D Face Shapes?

*Tal Hassner (Facebook AI, California, US)*



This talk aims to challenge long held and widely popular best practices for designing digital face processing pipelines. Specifically, nearly all face processing pipelines begin with face detection. Many systems then continue by localizing 2D facial landmarks for each detected face, a step typically used for face alignment and often also for 3D face reconstruction. Finally, depending on the application, some systems also estimate the 3D shape of each face. My talk proposes that these steps may be remnants of legacy designs from a time before effective deep learning was available, and are no longer required for many practical use cases. In fact, not only are these steps redundant, they also add unnecessary compute while introducing noise. As alternatives to these steps, I will share memory and compute efficient solutions for face detection, face alignment, and 3D face rendering. **The talk represents work done in academia, prior to joining Facebook / Meta AI and so does not represent that company in any way.*

## 3.14 On Implicit Avatars, Racial Bias, and the Light/Albedo Ambiguity

*Victoria Fernández Abrevaya (MPI für Intelligente Systeme – Tübingen, DE)*


**Joint work of** Victoria Fernández Abrevaya, Yufeng Zheng, Marcel C. Bühler, Xu Chen, Michael J. Black, Otmar Hilliges

We discuss two works covering two different aspects of 3DMMs.

In the first part we present IMAvatar [1], a new method for learning implicit morphable head avatars from videos. Traditional morphable face models provide fine-grained control over expression and pose, but cannot easily capture geometric and appearance details. Neural volumetric representations approach photorealism but are hard to animate, and do not generalize well to unseen expressions. To address this gap we introduce IMAvatar, a novel method for learning head avatars with an implicit representation, directly from monocular videos. Inspired by conventional 3DMMs, IMAvatar represents the expression- and pose-related deformations via learned blendshapes and skinning fields. We employ ray tracing and iterative root-finding to locate the canonical surface intersection for each pixel. The experimental results show that our method improves geometry and covers a more complete expression space, compared to state-of-the-art methods.

In the second part we shift the focus to face appearance. We find that current diffuse albedo estimation methods are biased towards light skin tones due to (1) strongly biased priors that prefer light skin, and (2) algorithmic approaches that do not address the light/albedo ambiguity. We discuss here a solution for the latter that builds on a key observation: the full scene image contains important information about lighting that can be used for disambiguation. Our experimental results show significant improvement compared to state-of-the-art methods on albedo estimation, both in terms of accuracy as well as fairness.

**References**

1 Y. Zheng, V. F. Abrevaya, M. C. Bühler, X. Chen, M. J. Black, O. Hilliges, *I M Avatar: Implicit Morphable Head Avatars from Videos.* arXiv, 2021.

# 4 Working groups

## 4.1 What are the 'Killer Applications' of 3D Implicit Representations

*Ben Mildenhall (Google Research – London, UK)*



Our first discussion topic, proposed by Ben Mildenhall, asked 'what are the 'killer applications' of 3D implicit representations?'. It was thought that applications in 3D vision are progressing slowly compared to 2D and that perhaps this was due to not yet currently having a killer application. Some felt that judging the slow development of 3D compared to current 2D technology was unfair due to 2D having more than 100 years of development and use in both technology and culture. Potentially, the slow development of 3D is similar to the original slow development of 2D over the past 100 years. It was agreed that we need to convince the general population that 3D computer vision is valuable and that cost of 3D capture technology is a significant factor. The only number the market understands is cost.

It was suggested that whilst one view is on what is tech feasible. Another discussion point is to imagine every consumer with this technology and what they would do with it? What universal adoption would look like? One interesting thing is the tendency of technology and games toward crafting a space to share with others. Animal Crossing, where users build and share islands or VR Chat, is a social VR game with no other objective than creating, sharing, and communicating. These have minimal to no game mechanics, just sharing. People could make an artistically pleasing space where others can gather in Augmented Reality (AR). Others suggested that some of the most exciting applications will be the democratisation of content creation. People either have 3D content creation skills or must go to a special effects company and ask for the content to be created for a price. However, neural rendering will enable massive reductions in the cost and skill required to create models. A simple spoken query to the network and a model will be generated for you.

Some felt that the current limiting factor in creating these experiences is difficulty in controlling or parameterising implicit representations. The killer applications enable users to control the neural representation, ask the game to look the way they like, and the network renders it for them. Along a similar line, it was generally agreed that we need to design and build improved ways of interacting with 3D content easily. 3D content on a 2D interface is not taking full advantage. In 2D, it is difficult for most users to interact with 3D content it is the user interfaces at the moment are challenging for many people. Similar to when 3D printers first arrived, it was the poor user experience that prevented widespread adoption.

Overall it was agreed by all that the future of 3D vision and neural rendering is inspiring and that the next decade will see an explosion of creative and exciting uses of these technologies. There are numerous directions to explore, but the combined progress of research and engineering in hardware and software will enable widespread adoption.

## 4.2 Are Neural Implicit Representations the Future of Morphable Models?

*Christian Theobalt (MPI für Informatik – Saarbrücken, DE)*



In the second flexible session, we tested a different format. Splitting into three groups, two physical and one remote, each group discussed the proposed topic for 30 minutes. We then regrouped and shared the results of the discussions with each other.

We concluded that in recent years, the research community had been excited to see that the implicit representation-based methods are achieving good performance in many aspects and are addressing many problems of morphable models, for example, in dealing with self-occlusion problems. They improve the rendering quality of scenes, faces, and objects and are also more efficient in the sense that only several images from different views are required for training. The trained network now being able to interpolate pose and generate images of almost the same quality. Compared to the morphable models using meshes, these methods do not need dense correspondences for supervision. Some methods do not even require information regarding the pose, volume or camera parameters.

On the other hand, we encourage further explorations of the generative ability of the implicit representation-based methods. We know that many questions for the morphable models or the mesh renderers still hold for the implicit models. For example, to build models with good generative ability, achieve relighting, render different materials, separate the albedo with shadow and highlight without losing any details, etc. It is also important to think about having smooth control over these properties. The morphable models solve this with dense correspondences, and the deformation space is well modelled so that we can have parametric control over specific properties. We encourage further thinking on how such correspondences should be encoded and how we can have smooth control over the deformations for the implicit models. Combining the Generative Adversarial Networks (GANs) is promising in this direction, yet we also expect other researchers to explore if a statistical model based on implicit representations is feasible or not. Implicit representations solve many problems that the morphable models have, yet we encourage further thinking on how to reach the same generative ability and smooth control over certain properties.

## 4.3 How Much 3D is Needed?

*Duygu Ceylan (Adobe Research – London, GB)*



In this workshop session, several questions were proposed for discussion. What is the best representation? Are 2D workflows good enough? Do we need explicit 3D? Is course 3D good enough? How much 3D do we need for 2D synthesis? Furthermore, how do we scale 3D methods to "in-the-wild" robustly? The group felt this was an old discussion that had been going on for many years.

The choice of needing 3D over 2D is very much task-dependent. Traditionally cartoons were created in 2D. Now, they are 3D even though they sometimes represent 2D as it is a more accessible representation to work with. Arguing over the coordinate system should be decided by the problem. For example, people used to argue over what edge detection is the best, but it was shown that all edge detections could be viewed as the same algorithm and depend on the task. It depends on the objective function. There is no right coordinate system for all problems. Depending on the cost function, if the cost function enforces 3D consistency, it was suggested that networks would learn to use 3D implicitly and learn from data as its the lowest energy state. Generally, it was agreed that one could do everything using just 2D, but 3D might need orders of magnitude less data, and it will be easier to learn.

The discussion amongst the group naturally flowed into a debate about how much humans learn from data and how much we use prior knowledge. For a lot of things we are now approximating using neural networks, we have physical theories; could we find these theories within the networks. Alternatively, could we use the networks to infer new theories about physical systems. It was argued that we would not be able to find physical laws from the weights of networks, but perhaps discovering laws that govern human behaviour is possible. The core issue is having the network communicate this information to us. For example, something like MuZero can play Go at a very high level, but we cannot understand the meaning behind its actions.

## 4.4 Metrical 3D Reconstruction of the Human Face

*Justus Thies (MPI für Intelligente Systeme – Tübingen, DE)*



Recent publications argue that a 3D reconstruction is need for various applications in AR/VR. These applications often rely on a metrically plausible reconstruction, since the face/human is displayed in a metrical environment (i.e., objects have a known scale). However, benchmarks for 3D face reconstruction like NoW [1] use a scale-invariant evaluation scheme (they search for an optimal scale to align the prediction and the reference). Measuring actual metrical reconstruction errors (i.e., only allowing for a rigid alignment without scaling) leads to a significantly different benchmark results and rankings. This discrepancy has to be discussed more prominently in the literature, since ideally we like to reconstruct metrical faces.

### References

1 S. Sanyal, T. Bolkart, H. Feng, M. Black, *Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision*, Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.

## 4.5 Synthetic Data Generation Using Morphable Models

*Federica Bogo (Meta Reality Labs Research – Zürich, CH)*



Morphable models have been successfully used to generate synthetic data. Recent examples are the FaceSynthetics [1] and AGORA [2] datasets. These works show how synthetic data generation is a flexible, powerful tool to train machine learning models, which can be used on real image data captured in the wild. This session focused on analysing the strengths and limitations of current synthetic pipelines and identifying how we can better leverage morphable models to generate higher-quality data.

It was suggested that an important question is how much one should focus on photo-realism. Minor discrepancies in local pixel intensity statistics between generated images and real ones can severely harm inference. However, rendering all the components realistically in an image (people, their interactions with the scene, background etc.) is very challenging in practice; this suggests one might need to invest significant effort (including manual work from technical artists) to obtain high-quality data.

Moreover, even with high-quality synthetic data, we might not be able to adequately capture the long-tailed distributions commonly found in the real world. A potential way to identify limitations in the data could be to enable a feedback loop from the trained model, tested on new examples, back to the training data itself to highlight problematic cases. Currently, the research community lacks the tooling to generate realistic synthetic data at scale quickly. We see some efforts in this direction [3], but there is still a lot to do for humans.

In the end, morphable models can play a crucial role in enabling a "virtuous cycle" from data collection to inference – which allows us to understand the world through our models – to synthetic data generation: better inference can enable the generation of higher-quality data; in turn, higher-quality generated data can enable the development of more accurate and robust models to perceive the world.

### References

1 E. Wood et al. *Fake it till you make it: Face analysis in the wild using synthetic data alone.* ICCV 2021.
2 P. Patel et al. *AGORA: Avatars in geography optimized for regression analysis.* CVPR 2021.
3 K. Greff et al. *Kubric: A scalable dataset generator.* CVPR 2022.

## 4.6 Morphable Models from Physics

*Dan Casas (Universidad Rey Juan Carlos – Madrid, ES)*



State-of-the-art data-driven approaches to model 3D garment deformations are trained using supervised strategies that require large datasets, usually obtained by expensive physics-based simulation methods or professional multi-camera capture setups. An alternative is to use physics-based deformation models. Formulating the problem as a set of physics-based loss terms that can be used to train neural networks without precomputing ground-truth data. Can this approach be applied to more areas of morphable models and computer vision in general?

The group liked this idea seeing it as a variant of thinking about what explicit real-world knowledge we have that can be used to simplify the situation. However, it gets more complicated when attempting to include all other physical properties, such as material parameters and friction in garment modelling. No physical simulation can do that at the moment; parameters of materials that are entirely unknown, the mass of yarns, for example, and other material properties are not fully modelled. This massive gap between the real and simulation world limits the application of these approaches. This simulation gap resulted in the suggestion that these models should really be called 'physics inspired' models as you are making a physical assumption about the world in the simulation, using a subspace of the physics.

Physics models are simplifications, so they are limited to the space that the model can represent; perhaps combining this first-principles approach with a data-driven refinement phase could be a way to go. Any explicit knowledge we can give the model aids in finding solutions with a good prior who needs the data. These approaches could help with explainability, as it is now possible to see what the model is understanding and doing to produce the effects.

Some expressed concern over how we combine physics models with neural scene representations. Understanding physics could help generalisation but the group see a challenge in combining these two. One of the benefits of implicit neural representations is that it can be done without correspondence.

## 4.7 Ethics

*Bernhard Egger (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)*
*William Smith (University of York, GB)*
*Christian Theobalt (MPI für Informatik – Saarbrücken, DE)*
*Stefanie Wuhrer (INRIA – Grenoble, FR)*



Our final workshop session was reserved for an extended discussion focusing on ethics. So often, as researchers, we are focused on new functionalities and challenges. However, what we are starting to develop now, for example, technologies that identify an individual's emotions, actions and identities, whilst enabling great opportunities, can have use cases that are extremely negative and not in the broader social interest. As leaders in this field, we know more about the possible issues than the general public and politicians. It is our responsibility to identify and predict any negative use cases and conceive of possible options to address these. Predicting negative use cases is a challenging problem, but one we need to have some answers for. To begin to address this, we split this discussion into four sections:

- Threats and ethical concerns.
- What should we do? What is our role?
- What more can our institutions do?
- Essential questions to ask ourselves when starting a project.

## Threats and Ethical Concerns

It was generally agreed that advances in neural rendering whilst having enormous positive use cases could enable some very concerning applications. The rise of Deep Fakes has shown the power of these approaches at producing misinformation, and the level of skill required to use these methods to create fake content is decreasing yearly. Further advances in these approaches to 3D computer vision will only enhance their realism and increase the difficulty of identifying real from fake. Similar unethical use cases involve any non-consensual use of a person's likeness; for example, in pornographic applications or as impersonation when using a telepresence system, it might not be possible to be confident that the person speaking is that person. Whilst privacy concerns over facial recognition systems are well documented. Morphable models are likely to be used to analyse people, emotions, attitudes and intents.

The group also discussed bias in the datasets and models we are developing. All datasets and models are biased, some to the point that the tools they form are potentially useless, offensive or upsetting. Do we try to understand the biases, or must we compensate for the biases somehow? In terms of facial recognition, it is now potentially the most important criterion, more so than accuracy. Research into ways to fix or quantify the bias is essential. However, it was also raised that it is sometimes a conscious decision to increase bias; for example, removing children from datasets might be the right thing to do. Ethical bias is not just due to data bias; an unbiased dataset can still have a model or data that has intrinsic bias. For example, women's faces might naturally have less variation in appearance, making recognition more challenging.

## What should we do? What is our role?

The question was raised as to whether we should consider banning a publication. Is it about preventing the research or preventing the use case? This is also challenging as there are many suitable applications of these technologies. It was generally agreed that it is likely we can not keep these algorithms out of the public sphere. That focus should be exclusively on public education and detection and preventative measures. There will always be bad actors, and it will be a cat and mouse game to combat them; we should be focusing on being ahead of these actors, not trying to prevent them from occurring, as that is impossible. Therefore, should we explicitly be working on counter models for the harmful applications of the technology we are developing? This could be an explicit part of the papers and research we do, not just considering.

Some felt that it would be necessary to create an algorithm equivalent to the FDA (Food and Drug Administration) that monitors public use cases of algorithms. Any algorithm released to the public would have to go through this administration. It could also be the focus of such an institution to educate the public. As researchers, we must push policymakers to address issues in these publications and institutions. We should be attempting to have regular meetings with policymakers that involve conversations about the impacts (positive and negative) and keeping people informed. Higher frequency of communication between researchers, the public and policy makers. Many of us have not had an opportunity to speak to policymakers. As academics, we are given by society many resources to do amazing things, and we must give something back. It is up to us to create these connections and start these dialogues. One can gain much from speaking with these groups and discussing these topics outside of one's research group.

## What more can our institutions do?

We discussed the different rules for different publications, i.e. CVPR requires approval from an institutional ethical review board, and NeurIPS asks reviewers to flag papers that concern them. Many felt that these rules were not sufficient to self enforce us as a community to think about these aspects. Some think it has to be on the reviewer's side, as asking the others to police themselves is not a practical task. Should it potentially be an ethical review board rather than the responsibility of the reviewers? Alternatively, should there be a multi-disciplinary and consistent board of scientists operating across conferences and checking flagged papers? Others felt ethical statements required by these papers were too generic to the point they are not helpful, raising concerns that if researchers are only thinking about the ethical question only at the point of writing the paper or at the end of writing the paper, it is already too late. Potentially we should be educating every researcher as a reviewer and give training on ethical considerations such that during the research, they are able to think more deeply about these issues.

## Essential questions before starting a project

Whilst the three previous discussions were fascinating and beneficial we felt it would be of use, especially for junior scientists, to collect a list of questions enabling an actionable output to support researchers in the field. Here we split into five groups, and each group was tasked with thinking of a set of questions one should ask themselves before starting a research project. These were then collated into a single document that could be a helpful starting guide for members of the seminar in future projects. The collated questions are listed below:

- What are potential misuses?
- What companies/institutions might be interested as well?
- What's the worst use case?
- Are people suffering from your research?
- What is the dataset you need to ensure there is no bias in your research? Is bias important for your research/use case?
- Would it be harmful in a democratic country?
- Who is funding the project?
- What is the field of application? Medical, entertainment, military?
- Privacy and consent of the data you would require or collect?
- Are people negatively affected by data collection? (Categorising, terrorist content online) Or obtained via nefarious means?
- Would any contracts or work provided by the project be ethical, provide a good standard of living?
- Will the world be a better place if this research is done?
- Am I using data for the purpose it was intended?
- When we pay participants to have their face or bodies captured, do they truly understand what they are giving away? Do they really understand how it will be used and how it could be used in future?

## Participants

- Thabo Beeler
Google – Zürich, CH
- Volker Blanz
Universität Siegen, DE
- Timo Bolkart
MPI für Intelligente Systeme – Tübingen, DE
- Dan Casas
Universidad Rey Juan Carlos – Móstoles, ES
- Bernhard Egger
Friedrich-Alexander-Universität Erlangen-Nürnberg, DE
- Victoria Fernandez Abrevaya
MPI für Intelligente Systeme – Tübingen, DE
- James Gardner
University of York, GB
- Oshri Halimi
Technion – Haifa, IL
- Patrik Huber
University of York, GB
- Marilyn Keller
MPI für Intelligente Systeme – Tübingen, DE
- Ron Kimmel
Technion – Haifa, IL
- Adam Kortylewski
MPI für Informatik – Saarbrücken, DE
- Chunlu Li
Universität Basel, CH
- Ben Mildenhall
Google – London, GB
- Nick Pears
University of York, GB
- Sami Romdhani
IDEMIA, FR
- Vincent Sitzmann
MIT – Cambridge, US
- Ayush Tewari
MIT – Cambridge, US
- Christian Theobalt
MPI für Informatik – Saarbrücken, DE
- Roy Velich
Technion – Haifa, IL
- Thomas Vetter
Universität Basel, CH
- Vanessa Wirth
Universität Erlangen-Nürnberg, DE
- Stefanie Wuhrer
INRIA – Grenoble, FR
- Tarun Yenamandra
TU München, DE

## Remote Participants


- Michael J. Black
MPI für Intelligente Systeme – Tübingen, DE
- Federica Bogo
Meta Reality Labs Research – Zürich, CH
- Duygu Ceylan
Adobe Research – London, GB
- Andrew Fitzgibbon
Graphcore – Cambridge, GB
- Tal Hassner
Facebook – Menlo Park, US
- Hao Li
Pinscreen – Los Angeles and UC Berkeley, US
- Xiaoming Liu
Michigan State University – East Lansing, US
- Arianna Rampini
Sapienza University of Rome, IT
- Emanuele Rodolà
Sapienza University of Rome, IT
- Shunsuke Saito
Reality Labs – Pittsburgh, US
- William Smith
University of York, GB
- Justus Thies
MPI für Intelligente Systeme – Tübingen, DE
- Maximilian Weiherer
Universität Erlangen-Nürnberg, DE
- Jiajun Wu
Stanford University, US
- Stefanos Zafeiriou
Imperial College London, GB

Report from Dagstuhl Seminar 22131

# Framing in Communication: From Theories to Computation

**Katarzyna Budzynska*[1], Chris Reed*[2], Manfred Stede*[3], Benno Stein*[4], and Zhang He†[5]**

1 **Warsaw University of Technology, PL.** katarzyna.budzynska@pw.edu.pl
2 **University of Dundee, GB.** c.a.reed@dundee.ac.uk
3 **Universität Potsdam, DE.** stede@uni-potsdam.de
4 **Bauhaus-Universität Weimar, DE.** benno.stein@uni-weimar.de
5 **Warsaw University of Technology, PL.** he.zhang@pw.edu.pl

## Abstract

Framing has become recognised as a powerful communication strategy for winning debates and shaping opinions and decisions. Entman defines framing as an action of selecting "some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described". Instead of engaging in costly and difficult exchanges of argument and counter-argument, a politician or a journalist can then try to reframe a dialogue on, for example, fracking from economic benefits to environmental hazards, or a dialogue on abortion from pro-life to pro-choice. Introduced in 1960's sociology, framing has been imported into communication sciences and media studies as an attempt to address the ways in which news is reported and, thus, a way in which to tackle manipulation and fake news. The topic has spread to other disciplines such as psychology, philosophy, semantics, pragmatics, political science, journalism, and, most recently – to computational linguistics and artificial intelligence. This seminar aims to pave the way to synthesising definitions developed in these theoretically and empirically driven areas and then to operationalise them in computational and applied areas by means of cross-disciplinary hands-on exchanges in facilitated discussions. Our goal is to support the development of innovative technologies, which can help us to quantify framing phenomena, to study framing at scale, and to deploy computational techniques in order to intervene against malicious attempts to influence opinions and decisions of the general public.

**Seminar** March 27–April 1, 2022 – http://www.dagstuhl.de/22131
**2012 ACM Subject Classification** Computing methodologies → Discourse, dialogue and pragmatics
**Keywords and phrases** Communication Strategies, Discourse and Dialogue, Computational Argumentation, Natural Language Processing
**Digital Object Identifier** 10.4230/DagRep.12.3.117

*Editor / Organizer
†Editorial Assistant / Collector


Framing in Communication: From Theories to Computation, *Dagstuhl Reports*, Vol. 12, Issue 3, pp. 117–140
Editors: Katarzyna Budzynska, Chris Reed, Manfred Stede, and Benno Stein
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

# 1 Executive Summary

*Katarzyna Budzynska (Warsaw University of Technology, PL)*
*Chris Reed (University of Dundee, GB)*
*Manfred Stede (Universität Potsdam, DE)*
*Benno Stein (Bauhaus-Universität Weimar, DE)*



Language is used for many purposes, both private and public. When speech or text is directed to wide audiences, it often aims at influencing stances, opinions, and dispositions of readers. This can be done by relatively transparent, rational argumentation, but also in considerably more subtle ways, by phrasing utterances in such a way that the underlying intent is noticed by readers more in passing – or not consciously at all. This is the realm of "framing", which concerns the careful selecting of the aspects of an event to be reported (those that fit the goal of letting a positive or negative evaluation shine through); the choice of terms that carry an inherent evaluation (e.g., "the frugal four" versus "the stingy four" in recent EU negotiations); and employing stylistic devices that correspondingly support the purpose (e.g., a monotonic versus a lively rhythm). Framing has been studied for quite some time, from many different perspectives, and it has also been covered by popular science books. Under these circumstances, it is not surprising that definitions and emphasis differ quite a bit between and even within disciplines – the notion of framing can itself be framed, too.

The computational research on language processing has addressed some of the linguistic purposes mentioned above: Sentiment analysis and opinion mining are well-established fields; argumentation mining has more recently caught much attention and is in the process of "settling down". Framing, being less transparent at the linguistic surface, has seen only very few attempts at formal modelling so far. The proposers of this seminar are convinced, however, that a computational treatment of framing is a central next step – extending opinion and argument analysis – and its operationalization calls for a deeper understanding of the term and the underlying mechanisms. Before computational theories can be formulated and applications be built, the potential contributions by the various relevant disciplines (sociology, political science, psychology, communication science, and others) should be studied carefully and assessed for potential common ground. This is the first purpose of the proposed seminar, and the second is the follow-up step of developing a roadmap for productive computational research toward the automatic identification of framing in text and speech, and modelling the connection to the underlying reasoning processes. To accomplish this, the seminar will address a relatively broad range of topics, covering relevant subfields of linguistics, computational modelling and application, as well as practical investigation of framing in the social sciences.

Framing, being less transparent at the linguistic surface, has seen only very few attempts on formal modelling so far. The proposers of this seminar are convinced, however, that a computational treatment of framing is a central next step – extending opinion and argument analysis – and its operationalisation calls for a deeper understanding of the term and the underlying mechanisms. Before computational theories can be formulated and applications be built, the potential contributions by the various relevant disciplines (sociology, political science, psychology, communication science, and others) should be studied carefully and assessed for potential common ground. This is the first area of the proposed seminar, and the second is the follow-up step of developing a roadmap for productive computational research toward the automatic identification of framing in text and speech, and modelling the connection to the underlying reasoning processes.

To accomplish this, the seminar addressed a range of topics, including:

- Argumentation theory, discourse analysis, rhetoric
- Journalism, political science, communication science
- Sociolinguistics, psycholinguistics
- Computational pragmatics and discourse modelling
- Computational social science and social media
- Computational models of argument and debating technologies

## 2 Table of Contents

**Executive Summary**
*Katarzyna Budzynska, Chris Reed, Manfred Stede, and Benno Stein* . . . . . . . . . 118

**Introductory talks**

Framing in Practice: Towards Computational Approach
*Konrad Kiljan* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121

Framing in Communication: From Theories to Computation Background: Discourse Analysis
*Andrea Rocci* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121

Framing in the Communication Discipline
*Jean Goodwin, Andrew Binder* . . . . . . . . . . . . . . . . . . . . . . . . . . . . 122

Computational Argumentation
*Henning Wachsmuth* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 123

Knowledge in Computational Argumentation
*Anette Frank* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 123

**Flash talks**

NLP Methods for Indoctrination Detection in German History Textbooks
*Lucie Flek, Dani Sandu, Ivan Habernal, Christopher Klamm, and Lars Wolf* . . . . 124

Detect – Verify – Communicate: Fact-Checking and Framing?
*Iryna Gurevych* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124

Framing(framing(framing(...)))
*Arno Simons* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125

Quantifying Luhmann: A Semi-Supervised Approach to Automatic Detection of Social Systems
*Martin Potthast* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125

**Working Groups**

Grounding and Theory: A Process-Oriented Approach to Framing
*Maud Oostindie, Anette Frank, Konrad Kiljan, Marcin Koszowy, John Parkinson, Andrea Rocci, and Joanna Skolimowska* . . . . . . . . . . . . . . . . . . . . . . . 126

Towards an Account of the Dynamics of Framing
*Zlata Kikteva, Annette Hautli-Janisz, Christopher Klamm, Steve Oswald, Chris Reed, Wassilik Siskou, Manfred Stede, and Henning Wachsmuth* . . . . . . . . . . . 129

Developing Benchmark Datasets for Frame Identification
*Khalid Al-Khatib, Katarzyna Budzynska, Alaxander Bondarenko, Lucie Flek, Anette Frank, Iryna Gurevych, Noam Slonim* . . . . . . . . . . . . . . . . . . . . . . . . . 133

Towards Operationalizing Frames through Axiomatization
*Lisbeth Allein, Philipp Cimiano, Ivan Habernal, Martin Potthast, Arno Simons, and Benno Stein* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136

**Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 140

# 3 Introductory talks

## 3.1 Framing in Practice: Towards Computational Approach

*Konrad Kiljan (University of Warsaw, PL)*



Despite its ambiguity, the notion of framing has been used for decades in various trainings aimed at boosting speakers' persuasiveness and communication skills. Wide recognition of framing as an extremely effective tool in media studies and debate education resulted in it becoming an umbrella term applied to multiple techniques for marketing purposes. This talk proposes a reduction in the term's scope with the aim of covering by it only the aspects recognised across both domains. A context-weary content analysis can then be applied to categorise framing attempts in accordance with Habermas's classification. The second part of the session included a set of practical exercises allowing the seminar's participants to reflect on the lived experience of framing in communication to enrich their sensitivity to the implicit notions that are often difficult to map while analysing transcripts.

## 3.2 Framing in Communication: From Theories to Computation Background: Discourse Analysis

*Andrea Rocci (University of Lugano, CH)*



This introductive presentation was aimed to show how a classic notion of frame derived from linguistics semantics can serve as an operational concept in discourse analysis ([8],[7]). Communication scholars are familiar with Goffman's notion of frame as a basic definition of a situation "built up in accordance with principles of organization" that shape the understanding of events and regulate social events and "subjective involvement" in them [4]. This famously non-operational notion has formed the basis of various attempts at "frame analysis" aimed at reconstructing culturally shared patterns of interpretation used by communicators. Most of these literature is however unaware of the parallel concept frame developed in linguistics ([1],[2],[3]), due primarily to the work of Charles Fillmore on Frame Semantics since the early 1970s. This frame notion emerges as a direct development of the concept of the argument frame of a predicate, including the roles (deep cases, theta-roles) that characterize each argument place. In fact, Frame Semantics shows that the meaning of lexical predicates has to be understood relative to largely tacit, structured background scenes or frames. Thus, the linguistic and Goffmanian notions of frame are reconciled. From the point of view of the rhetorical choices of the communicator, framing involves two levels of meaningful choice. At a first level, the communicator can decide to present a given situation according to different conceptual frames. A classic example of alternative framing is offered by Aristotle in Rhetoric (III, 2, 1405b) when he observes that the Orestes can be rightly called both **mother-slayer** and **father's avenger**. The two epithets select alternative framings of the very same action perpetrated by Orestes. Both frames can be truthfully predicated of the situation, but their evaluative implications are opposite.

At a second level, once a given frame has been chosen, the choice of the specific lexical predicates and syntactic construction within it can serve to selectively activate certain components of the frame and to select a viewpoint on the scene.

Tropes such as metaphor and metonymy allow, respectively, the cross-domain mapping of frame structure and the collapsing of distant but related scenes into a unitary humanly perceivable frame ([5],[9]).

The cases of framing considered up to this point, both literal and metaphorical/metonymical, concern the propositional content of the utterance. Yet, framing can be applied also at the pragmatic level of utterances. Pragmatic frames are not different in kind from semantic ones, much like performative verbs are not really different from other lexical predicates in most respects, including the fact they have an argument frame defining a series of roles [6].

**References**

1 C. J. Fillmore, "Lexical semantics and text semantics," *New directions in linguistics and semiotics*, vol. 32, pp. 123–147, 1984.
2 C. J. Fillmore, *Form and meaning in language*, vol. 1. CSLI Publications, Center for the Study of Language and Information, 2003.
3 C. J. Fillmore and C. F. Baker, "A frames approach to semantic analysis," in *The Oxford handbook of linguistic analysis*, pp. 313–339, Oxford University Press, 2012.
4 E. Goffman, *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.
5 C. Pollaroli and A. Rocci, "The argumentative relevance of pictorial and multimodal metaphor in advertising," *Journal of argumentation in context*, vol. 4, no. 2, pp. 158–199, 2015.
6 A. Rocci, "Manoeuvring with voices the polyphonic framing of arguments," *Examining argumentation in context: Fifteen studies on strategic manoeuvring*, vol. 1, p. 257, 2009.
7 K. B. Jensen and R. T. Craig, *The International Encyclopedia of Communication Theory and Philosophy, 4 Volume Set.* John Wiley & Sons, 2016.
8 A. Rocci and M. Luciani, "Semantics and verbal communication," *Verbal communication*, pp. 57–76, 2016.
9 A. Rocci, S. Mazzali-Lurati, and C. Pollaroli, "The argumentative and rhetorical function of multimodal metonymy," *Semiotica*, vol. 2018, no. 220, pp. 123–153, 2018.

## 3.3 Framing in the Communication Discipline

*Jean Goodwin (NC State University)*
*Andrew Binder (NC State University)*



Research on framing in Communication emerged in the 1970s, influenced by work on cognitive frames (social psychology), situational frames (sociology) and interactional frames (discourse analysis). From the Communication perspective, all messages are framed: communicators select the information they convey and present it in a way that makes some aspects more salient, others less. On some topics, such as science-based issues, lists of typically deployed news frames have been developed, but little attention has yet been given to identifying "master" frames across topics. And as Scheufele has cautioned, it is a mistake to assume that a communicator's (discursive) message framing straightforwardly induces the audience to adopt a (cognitive) framing. The evidence for the relatively small persuasive effects of framing emerges predominantly from studies of equivalency framing, in which the same information (how much water is in the glass) is conveyed in different manners ("half full/half empty"). In these cases, the context provided by the message may induce the audience to apply a cognitive scheme, coming to see the topic as that frame. But much framing of interest is emphasis framing, in which some aspects of a complex situation are made salient. Such framing can make an aspect more cognitively accessible, but is likely to have little persuasive effect, especially in an environment where there are numerous competing frames. Finally, interactional framing – the ways interlocutors make sense of their communication – remain understudied.

## 3.4 Computational Argumentation

*Henning Wachsmuth (Universität Paderborn, DE)*



Computational argumentation deals with computational analysis and synthesis of natural language arguments. In this tutorial talk, we give an overview of computational argumentation from a natural language processing (NLP) prospective. Starting from basics of human argumentation, we introduced the main argument mining, argument assessment, and argument tasks. We detail how to approach such tasks with NLP methods on the example of stance classification before we provide insights into the main applications of computational argumentation. On this basis, we discuss the relation of computational argumentation to framing in communication.

## 3.5 Knowledge in Computational Argumentation

*Anette Frank (Universität Heidelberg, DE)*



In this background talk on argumentation, I am stressing the need of knowledge in computational argumentation. I first point to the role of context vs. content in computational argument analysis, where we have shown that current algorithms for argument analysis are strongly relying on contextual signals, like discourse markers – at the cost of content. This can lead to undesirable model biases, especially when being confronted with novel task settings or data distributions.

I then demonstrate recent work conducted in the ExPLAIN project, which aims to reconstruct implicit background knowledge in natural language arguments – which is easy for humans to fill in by reading between the lines, but where computational systems struggle. We identify the relevance of commonsense knowledge and showcase that by including such knowledge resources in downstream computational argumentation tasks we can improve system performance. We then show that background knowledge a system uses to make such implicit knowledge explicit in arguments can be generated in natural languages – which helps to make the process transparent and controllable.

# 4 Flash talks

## 4.1 NLP Methods for Indoctrination Detection in German History Textbooks

*Lucie Flek (Universität Marburg, DE)*
*Ivan Habernal (TU Darmstadt, DE)*
*Christopher Klamm (Universität Mannheim, DE)*
*Dani Sandu (European University Institute, IT)*
*Lars Wolf (TU Darmstadt, DE)*


**Main reference** Lars Wolf: "NLP methods for indoctrination detection in German history textbooks" (Master Thesis), TU Darmstadt, 2021.

Controlling information and mass media is crucial for dictators to stay in power. While propaganda and fake news detection has seen a surge in research attention lately, this work focuses on analyzing deeper beliefs and values. As a collaboration between political science and computer science, we introduce the novel task of indoctrination detection. We processed 46 scanned textbooks from the German Democratic Republic (GDR) and the Federal Republic of Germany (FRG), used in history classes from 1948 to 1989 and covering the two countries' common history from 1900 to after World War II. We automatically analyze these textbooks regarding several facets of indoctrination, which include gatekeeping, selective attribution, subjective language, and appropriation. For examining these, we use embedding-, semantic role labeling- and emotion-based techniques to identify word meaning shifts, activity and passivity of entities and emotions towards entities in the textbooks. We then create a corpus for the new task of indoctrination detection by manually annotating 336 excerpts of the history textbooks for indoctrination mechanisms and entities affected. We use this new corpus to train a machine learning model for indoctrination detection, evaluating the predictive power of the semantic features we developed based on the insights we gained from our analysis. We demonstrate that the NLP techniques can mainly capture emotionally loaded expressions, while still struggling with broader subtle contexts.

## 4.2 Detect – Verify – Communicate: Fact-Checking and Framing?

*Iryna Gurevych (TU Darmstadt, DE)*



Combating misinformation is a challenge the information society approaches by equipping computer users with effective tools for identifying and debunking fake news. However, current Natural Language Processing (NLP) techniques are computationally expensive, fall short of fighting real-world misinformation, and do not adequately address real-life scenarios. Additionally, we believe automatic NLP systems should also communicate against misinformation in a manner persuasive to the end user. In this talk, we briefly discuss our ongoing work on these topics. Namely, we are pursuing research that addresses misinformation detection with systems that are more data efficient and less expensive. To narrow the gap between NLP and real-world fact-checking, we constructed two richly annotated fact-checking datasets using (i) real-world claims from Snopes and (ii) real-world-like claims from search queries with long documents. Finally, to edify false beliefs, we are collaborating with cognitive scientists and psychologists to create a system that automatically detects and responds to attitudes of vaccine hesitancy, encouraging anti-vaxxers to change their minds with effective communication strategies. These strategies work by affirming beliefs, reframing the anti-vaxxer argument to point out flawed logic, and hopefully dissuading someone from believing false information. This is a joint work by Iryna Gurevych, Andreas Hanselowski, Nils Reimers, Max Glockner, and Luke Bates.

## 4.3 Framing(framing(framing(...)))

*Arno Simons (DZWH – Berlin, DE)*



The computational treatment of framing presupposes a deep understanding of the term and the underlying mechanisms. To gain such an understanding, a historical view on the genesis and development of the framing concept is instructive. In his talk, Arno Simons traced the concept from its use today back to key works in the 1960s and even further to its roots in gestalt psychology, pragmatism, phenomenology, and early sociology. This historical mapping revealed that the framing concept has been intimately linked to the idea that our "reality" is socially constructed. Also, the term framing has been defined and used in distinctively different ways. It can refer to both processes and outcomes of processes, and it can focus on either the psychological or the sociological level, or both. When modelling framing computationally, we should be aware of and transparent about which definition of the concept we are following. Equally important, we must understand that the computational modelling of framing does not necessarily call for a completely new method or toolkit, because many existing tools, from named-entity recognition over topic modelling to argument mining already capture essential aspects of framing, at least in ways that we could harness if we paid attention. Finally, Arno argued that we should reflect on the ways in which our modelling of framing, e.g. in the form of fact-checking implementations, is itself a way of framing reality, which might feed back into the very social phenomena we are trying to serve or analyze with our applications and algorithms.

## 4.4 Quantifying Luhmann: A Semi-Supervised Approach to Automatic Detection of Social Systems

*Martin Potthast (Universität Leipzig, DE)*



Niklas Luhmann is a well-known German social systems theorist who proposed a functional differentiation of society. His theory has been widely recognized in the social sciences, but it is just one among many. In our work, we attempt for the first time to quantify the social systems theory according to Luhmann. We do so by harnessing the books of Luhmann himself: He wrote 8 books, one for each system he identified. The books, divided into passages, serve as a labeled ground truth for texts that pertain to a given social system. To develop a method to classify texts into social systems, we employ seed-guided text classification, where a number of seed words are derived from Luhmann's books that are discriminative of each individual system compared to the others. This is done with the goal of transfersing a trained model from the domain of Luhmann's books to more generic text domains, such as Wikipedia, news articles, or other scientific articles. Our approach shows promising results, indicating that a classification of text into social systems is indeed possible. This may give rise to quantitative analyses of social systems in social sciences, supporting social scientists in their daily work.

# 5 Working Groups

## 5.1 Grounding and Theory: A Process-Oriented Approach to Framing

*Maud Oostindie (Maastricht University, NL)*
*Anette Frank (Universität Heidelberg, DE)*
*Konrad Kiljan (University of Warsaw, PL)*
*Marcin Koszowy (Warsaw University of Technology, PL)*
*John Parkinson (Maastricht University, NL)*
*Andrea Rocci (University of Lugano, CH)*
*Joanna Skolimowska (Warsaw University of Technology, PL)*



### 5.1.1 Introduction

Framing, for Bateson [2] and Goffman [7], is a process through which people make sense of everyday situations and contexts by highlighting and/or excluding specific information. This process creates frames, which are in their turn challenged, and re-created through discursive practices. Myriad authors and disciplines have since engaged with the notion of framing, but the concept remains fuzzy, with different disciplines defining and using the concept in different ways. In his 1993 article, Entman [5] raises this issue and claims that "framing is often defined casually, with much left to an assumed tacit understanding of reader and researcher" (p. 52). This fuzziness as well as the casual and diverse nature of the engagement with the concept across academic disciplines has important implications for the empirical investigation and application of framing. If we do not have a precise understanding of framing, how are we able to recognize framing in the wild?

The aim of the Dagstuhl seminar on framing in communication was to make framing recognizable and understandable for empirical investigation. Specifically, the seminar aimed to make a start with making the step from framing theory to computation, creating guidelines for computational models of framing. The aim of our breakout group was to focus on the theory part of the seminar, and to develop a set of cues for empirical analysis that can eventually inform computational models. Starting from the premise that **framing is a recognizable phenomenon**, we set out to identify ways in which this phenomenon can be recognized manually. Our aim is to develop a minimal and operationalizable understanding of framing. Our aim here is not to provide an operationalization for quantitative or computational analysis of framing. We come back to this on the proposed research (cf. Section 5.1.4). First, we introduce a bit of relevant academic background, and then we outline the main findings of our breakout group during the Dagstuhl seminar.

### 5.1.2 Background

In the overlapping fields of communication, discourse analysis, and argumentation theory, some recent work has been done on:
1. theorizing framing and
2. employing novel methodologies in the study of (re)framing.

In terms of theoretical findings, framing has been associated with offering the audience a salient premise in a deliberative process that can ground decision and action Fairclough and Mădroane [6] or assumptions (basic perspectives) underlying the debate [8]. In terms of methods in the study of framing in communication, Aakhus and Musi [12] have employed frame semantics and knowledge-driven argument mining to retrieve semantic frames present in a large corpus about fracking. Specifically, the words associated with the core elements of a semantic frame have been automatically retrieved in order to map how different actors, positions, and venues of discussion are assembled around what is treated as irreconcilable in the controversy. A phenomenon of rephrasing which may be dealt with in terms of changing frames by the speakers has been recently studied in relation to argumentation structures [16], by employing corpus linguistics [10] and experimental [15] methods to study its persuasiveness.

On top of that, relevant work on framing has been done at the intersection of communication, deliberation, and conflict. Specifically, the authors investigate how people (re)frame situations and interactions in deliberative and informal encounters. Black [4] demonstrates how the method of storytelling gets used to reframe conflict. Since framing is a discursive process centred on communication, people can challenge and alter how interaction is framed. By reframing issues and relationships, people reframe the meaning of a conflict [4].

### 5.1.3 Findings

Applying methods and theories of anthropology, semiotics, argumentation and cultural studies, we forged a working definition of **framing as a shared scene building process**. This process-oriented understanding of the communicative practice of framing takes into account both its constitutive linguistic elements and the social contexts it refers to in order to trigger associations with them [1]. The rhetorical power of framing derives from the fact that it allows the participants of dialogues (and other types of communication) to convey and extract meaning from smaller bits of text in a way that is very economical, yet not necessarily stylistically coarse. As a result, single utterances evoke entire stories which convey concepts of purpose, value, efficacy and self-worth [3]. Being a means of building mutual understanding between the participants of communication, framing can be used for informative as well as persuasive aims. At the same time, framing is a dynamic process that helps people communicate and reach a mutual understanding. In this sense, framing is not just instrumental but also communicative [9].

The elements used to build frames can be categorized into three main types:
1. things (agents, entities),
2. properties (attributes) and
3. relationships.

As indicated by theatre practitioners [11], these basic components allow for the construction of scenes that are later read as building blocks of recognizable stories. Referring to broader stories gives a speaker a chance to make use of their rhetorical power without revealing them all. Successful framing exposes the listeners to a scene in a way that allows them to grasp the essential meaning of the entire story [13]. The way ingredients of a particular scene are interpreted and linked with a specific story depends on the individuals' familiarity with the socially acquired depository of their types and their sensitivity towards their indicators. As a single symbol, metaphor, storyline, perspective or stylistic cue can transform the entire meaning of a frame, it might often be misinterpreted or fail to succeed due to differences in the patterns of interpretation adopted by dialogue participants.

An exemplary usage of framing can be observed in Russian propaganda pieces describing the 2022 invasion of Ukraine as a "result of NATO expansion". In terms of actors, properties and relationships, it refers to stories and worldviews from the Cold War period in which only two powerful imperialist blocks are treated as decision-making agents. Leaving out an independent nation from the scene and describing one side as an expansive aggressor suggests that escalation of war should be read not even as a response, but a direct consequence of an equilibrium-seeking system. Countering this narrative in a dialogical situation would require the introduction of other actors, reinterpretation of their properties and relationship and can be successful only by reference to other, at least equally deeply embedded stories that the audience can later find more in line with other sources of knowledge and meaning.

Framing and reframing in communication, thus, relate to reframing the main aspects of frames: things (agents, entities), properties (attributes), and relationships. We term the reframing of things **compositional framing**, the reframing of properties **attributive framing**, and the reframing of relationships **relational framing**. This implies that the empirical cues for identifying frames are to identify the things in a frame, the properties, and the relational structure.

### 5.1.4 Proposed Research

Presented process-oriented approach to framing delivers a starting point to build computational models for mining and employing framing in natural language, however, a proper operationalization of the concept is needed. We propose taking a few steps back and theorizing first about a way for **qualitative analysis of framing**. The next step would be to develop a coding scheme for **quantitative analysis**. Only then, we believe, it is possible to think about computational models. For us, one of the main challenges in the aim of developing a computational model for framing is the complexity of the conceptualization.

**References**

1 Aminoff, Elissa M. "Putting Scenes in Context." Scene Vision: Making Sense of What We See, edited by Kestutis Kveraga and Moshe Bar, The MIT Press, 2014, pp. 135–54
2 Bateson, G. (1972). Steps to an ecology of mind. Ballantine Books.
3 Baumeister, Roy F. Meanings of Life. Guilford Press, 1991.
4 Black, L. W. (2020). Framing democracy and conflict through storytelling in deliberative groups. Journal of Deliberative Democracy, 9(1).
5 Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. McQuail's reader in mass communication theory, 390-397.
6 Fairclough, I. & Mădroane, I.D. (2016). An argumentative approach to policy 'framing'. Competing 'frames' and policy conflict in the Roşia Montană case. Rozenberg Quarterly
7 Goffman, E. (1974). Frame analysis: An essay on the organization of experience. Harvard University Press.
8 Goodwin, J. (2019). Re-framing climate controversy: The strategies of The Hartwell Paper. In Proceedings of the 9th Conference of the International Society for the Study of Argumentation.

9 Habermas, J. (1984). The theory of communicative action (T. McCarthy, Trans.). Beacon Press.
10 Konat, B., Budzynska, K., & Saint–Dizier, P. (2016). Rephrase in argument structure. In P. Saint Dizier & M. Stede (Eds.), Foundations of the Language of Argumentation. COMMA 2016 Workshop. University of Potsdam, September 13, 2016 (pp. 32–39). University of Potsdam.
11 Murphy, Vincent. Page to Stage: The Craft of Adaptation. University of Michigan Press, 2013
12 Musi, E. & Aakhus, M. (2019). Framing fracking: Semantic frames as meta-argumentative indicators for knowledge-driven argument mining of controversies. Journal of Argumentation in Context, 8(1), pp. 112- 135. doi: 10.1075/jaic.18016.mus
13 Park, Soojin, & Marvin M. Chun. "The Constructive Nature of Scene Perception." Scene Vision: Making Sense of What We See, edited by Kestutis Kveraga and Moshe Bar, The MIT Press, 2014, pp. 45–72
14 Putnam, L. L., & Holmer, M. (1992). Framing, reframing, and issue development.
15 Schumann, J., Zufferey, S., & Oswald, S (2019). What makes a straw man acceptable? Three experiments assessing linguistic factors. Journal of Pragmatics, 141(1), 1–15. https://doi.org/10.1016/j.pragma.2018.12.009
16 Visser, J., Koszowy, M., Konat, B., Budzynska, K., & Reed, C. (2018). Straw man as misuse of rephrase. In Steve Oswald and Didier Maillat (Eds.), Argumentation and Inference. Proceedings of the 2nd European Conference on Argumentation, Fribourg 2017 (pp. 941–962). College Publications.


## 5.2 Towards an Account of the Dynamics of Framing


*Zlata Kikteva (Universität Passau, DE)*
*Annette Hautli-Janisz (Universität Passau, DE)*
*Christopher Klamm (Universität Mannheim, DE)*
*Steve Oswald (University of Fribourg CH)*
*Chris Reed (University of Dundee, GB)*
*Wassiliki Siskou (Universität Konstanz, DE)*
*Manfred Stede (Universität Potsdam, DE)*
*Henning Wachsmuth (Universität Paderborn, DE)*




### 5.2.1 Introduction

Framing is a dynamic process that allows for the use of multiple frames within a single discourse unit. This presupposes that frames must undergo certain changes and/or interact with each other. The goal of our working group was to identify the types of interactions that happen between and within frames, so-called 'framing moves', and to examine the way they occur, i.e., patterns of framing dynamics.

While our working assumption is that similar framing dynamics can be found in both dialogues and monologues, we focused on the dialogical structures to narrow the scope of the research. We embed patterns of the framing dynamics in the dialogue protocols offered by [2] in order to put some constraints on them. Our account of framing dynamics is constrained by dialogue type, i.e., within an instance of a given type (say, a negotiation dialogue) each of the framing moves is available according to protocol and constraints that are proper to the dialogue type.

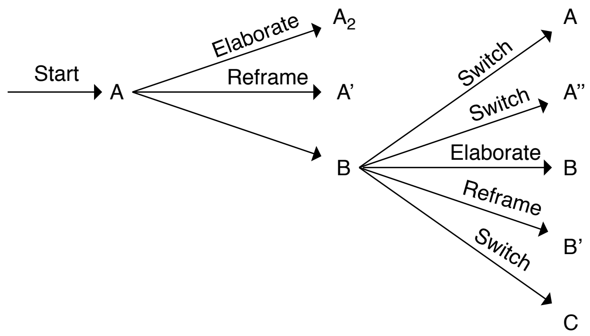



**Figure 1** An abstract example of different options of transitions for evolving an initial frame $A$ given the introduced framing moves as the edges $\xrightarrow{\text{move}}$ between frame $A$ and $B$ (e.g., $A \xrightarrow{\text{switch}} B$).

## 5.2.2 Research Progress

### 5.2.2.1 Frames in Context

As we view frames dynamically, we decided not to focus in detail on the definition of framing. Instead, we bracket it out with an underspecified placeholder definition into which a variety of more specific definitions might be slotted. For the purposes of this research, a frame is an assignment, either intentional or non-intentional of salience with respect to a set of information. It can be thought of as a vector, $\vec{F}$, that allocates a set of changes to salience values for an extensive set of information units (e.g., a set of relative percentages increases and decreases of salience values for all the propositions in a knowledge base). This simplistic model allows multiple frames to be applied simultaneously, allows frames to incorporate definitions that rest heavily on models of topics, and allows frames to be compared quantitatively and qualitatively. Frames apply to a unit of contiguous discourse material, called a Frame Unit (FU). Frame shifts occur between two adjacent FUs.

### 5.2.2.2 Patterns of Framing Dynamics

We can identify *seven* crucial components to assess the underlying dynamics. We extracted the dominant patterns of framing dynamics in an inductive manner using prototypical examples from various domains and call each pattern a 'framing move'. Below, we provide the definitions of each identified framing moves as a new framework for modelling framing dynamics:

- **Start**: *initiates or introduces a new (initial) frame*
- **Take on**: *accepts a frame and continues it; new speaker*
- **Elaborate**: *increases or reduces specificity of a frame; same or new speaker*
- **Reframe**: *modifies a frame but maintains some continuity; same or new speaker*
- **Switch**: *introduces a new (different) frame without necessarily rejecting the previous frame; same or new speaker*
- **Reject**: *rejects the suitability of a frame; new speaker*
- **Merge**: *selectively combines two or more frames; same or new speaker*

In Fig. 1, for example, we illustrate the interplay and transitions between different frame types. An initial frame $A$ may be elaborated into $A_2$ or reframed into $A'$. There may also be a switch to a frame $B$. From $B$, several further options evolve (as holds for $A_2$ and $A'$

but is not shown). We will illustrate these patterns in a real-life example with the following constructed dialogue excerpts on the topic of **crime rate** with two speakers:

SPEAKER 1 – **start**: *Crime is a dreadful plague in this country.* [*TOPIC: CRIME RATE | FRAME: ILLNESS/PLAGUE*]

SPEAKER 2 – **take on**: *Indeed, this infection needs to be eradicated.* [*TOPIC: CRIME RATE (INHERITED FROM PREVIOUS TURN) | STANCE: SAME | FRAME: ILLNESS (INHERITED FROM PREVIOUS TURN)*]

SPEAKER 2 – **elaborate**: *It's infecting our cities, our towns and our boroughs.* [*TOPIC: CRIME RATE (INHERITED FROM PREVIOUS TURN) | STANCE: SAME | FRAME: ILLNESS (INHERITED FROM PREVIOUS TURN)*]

SPEAKER 1 – **reframe**: *Yeah, it's like a cancerous tumour that just keeps on growing.* [*TOPIC: CRIME RATE (INHERITED FROM PREVIOUS TURN) | STANCE: SAME | FRAME: ILLNESS-CANCER*]

SPEAKER 1 – **switch**: J*ust look at the numbers [: the murder rate is up 10% per year, now over 100,000 homicides annually.]* [*TOPIC: CRIME RATE (INHERITED FROM PREVIOUS TURN) | STANCE: SAME | FRAME: STATISTICS (NEW)*]

SPEAKER 2 – **reject**: *Hey – consider how the citizens suffer from this constant threat of burglaries.* [*TOPIC: CRIME RATE (INHERITED; POSSIBLY SWITCHING TO SUBTOPIC) | STANCE: SAME | FRAME: POPULAR WELL-BEING*]

### 5.2.3 Data Analysis – Real-Life Dialogue Example

We applied the concepts of framing moves to a real-life example dialogue. We chose an excerpt previously analyzed in terms of blends (defined as integrated mental spaces recruiting conceptual input from different input spaces) by [1] to test the applicability of the framing moves. The dialogue is taken from Loveline, a call-in radio show in North America that gives listeners medical and relationship advice[1]. In the following dialogue, the hosts are reacting to the caller's concern about getting two orgasms in a row:

ADAM – **start**: Well listen, the Lord was kind to you that day. [*FRAME: MIRACLE*]

DR. DREW – **reframe**: He spoke directly to him. [*FRAME: MESSAGE*] [*Embed/push new dialogue type*]

ADAM – **start**: Drew, do you think anything's wrong with the guy? [*FRAME: HEALTH ISSUE*]

DR. DREW – **take on**: No, no, no [*FRAME: HEALTH ISSUE*]

ADAM: Well listen just enjoy it. [*Unembed/pop dialogue type*]

ADAM – **reframe**: It happened to you once. It'll be like sort of a Holy Grail you chase for the rest of your life. [*FRAME: GRAIL*]

ADAM – **elaborate**: But y'know count yourself among the blest. It happened to you once and that's more than it's happened to me.

DR. DREW – **reframe**: Well this could be some kind of Purgatory, [*FRAME: RELIGION*]

DR. DREW – **switch**: sort of a Sisyphus-like constantly trying to recreate that and never quite achieving it. [*FRAME: SISYPHUS*]

ADAM – **merge**: It is sort of a strange thing that you have this incredible sort of never-ending orgasm once and then end up chasing it like it was Moby Dick for the rest of your life. [*FRAME: MOBY DICK*]

[1] http://kroq.radio.com/shows/ (offline)

The example illustrates almost all of the identified framing moves. Adam begins the dialogue with the frame of a miracle which is then reframed by Dr. Drew who brings up the idea of a message. After the two humorous comments, Adam in a more serious manner starts a new frame of the health issue which is then taken on by Dr. Drew when he responds to the question. Adam goes back to the humorous approach with a reframe and elaboration when he talks about the experience of the caller as a Holy Grail after which Dr. Drew suggests an alternative view on the experience by calling it a purgatory thus reframing again. Dr. Drew switches the frame completely by comparing the caller, who will be trying to achieve something impossible, to Sisyphus. The dialogue excerpt is concluded with Adam merging the previously mentioned frames of the Holy Grail by referring to the caller's experience as something 'incredible' with the frame of Sisyphus ('you end up chasing') into a new frame of Moby Dick.

### 5.2.4 Future Directions

Understanding the framing dynamics in dialogues can help readers reflect on the (non) intentional use of framing at a structural level beyond the content itself and can help us understand the effect of the interplay between multiple frames in a text. Our preliminary work provides a starting point for a comprehensive study of the dynamics of framing. We presented seven crucial components to assess the underlying dynamics. These framing moves were extracted in an inductive manner using two prototypical examples. We plan to extend this process by conducting an annotation study on different datasets (e.g., politics, humour, etc.) with multiple annotators to create a robust scheme in order to test our presented framework on a large scale. Given the close connections to related subfields such as dialogue protocols, future work will further strengthen our proposed framework's unique position and novelty within the research landscape.

#### References

1 Seana Coulson. Extemporaneous blending: Conceptual integration in humorous discourse from talk radio, Style 39 (2) : 107-122, 2005
2 Douglas N. Walton & Erik C. W. Krabbe. Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. State University of New York Press, Albany NY, 1995

## 5.3 Developing Benchmark Datasets for Frame Identification

*Khalid Al-Khatib (University of Groningen, NL)*
*Katarzyna Budzynska (Warsaw University of Technology, PL)*
*Alexander Bondarenko (Martin-Luther-Universität Halle-Wittenberg, DE)*
*Lucie Flek (Universität Marburg, DE)*
*Anette Frank (Universität Heidelberg, DE)*
*Iryna Gurevych (TU Darmstadt, DE)*
*Noam Slonim (IBM – Haifa, IL)*



### 5.3.1 Introduction and Background

Constructing high-quality datasets for *frame identification* is an essential step in developing and evaluating corresponding computational models. To build a framing dataset, different questions can be asked such as: What are the possible typologies of frames? How to outline clear and practical annotation guidelines? And how to ensure the feasibility of the annotation given the complex nature of the task? Here, we discuss a pilot annotation study that seeks to deliver some preliminary answers to the questions above. We strive to propose frame typology in a relevant and appealing manner to the Natural Language Processing (NLP) community and back this up with real-world examples and a small dataset.

#### 5.3.1.1 Existing Framing Datasets

Card et al. developed the media frame dataset [3], conducting an annotation study of nearly 16,000 news articles on the topics of same-sex marriage, immigration, and smoking. Frame labels are adopted from Boydstun et al. set [2] that contains 15 general frames, including quality of life and public opinion, among others. In the computational argumentation area, Ajjour et al. modelled frame as a group of arguments that focus on a certain *aspect* such as "economics" [1]. The frame labels were derived from the debate portal "debatepedia.org", where users provide topic-specific aspects for arguments. The dataset comprises around 12,000 arguments belonging to 1,623 frames.

Heinisch and Cimiano [4] experimented with the Ajjour et al.'s [1] and the media frames Card et al. datasets [3]. Comparing the two studies regarding the used frame ontologies, Heinisch and Cimiano [4] emphasise three main facets of differences: (1) frame granularity (15 vs. 1,623), (2) domain of arguments (news articles on policy debates vs. any topic proposed by online users), and (3) ontology origin (experts annotation vs. online users meta-information).

### 5.3.2 Pilot Annotation Study

#### 5.3.2.1 Data Source

We picked up the recent controversial topic about transgender "Lea Thomas", a swimmer who won a national college championship in the United States. This event sparked an intense debate about whether it was fair for transgender female athletes to compete with Cis females. We selected three articles with different views on this topic using the web portal `www.allsides.com`. This portal publishes various articles on similar news stories from different political viewpoints: right, left, and centre (cf. Figure 2).

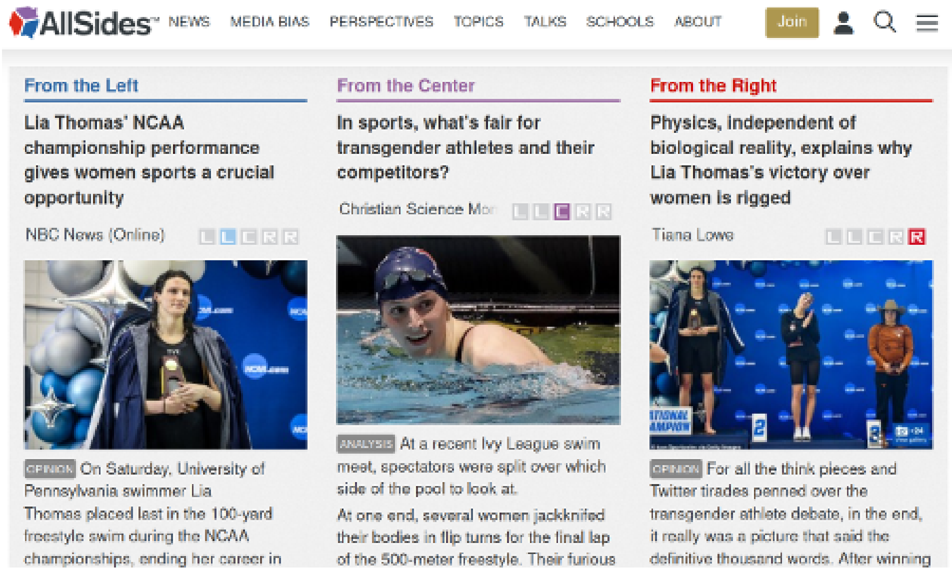



**Figure 2** Articles on AllSides about Lia Thomas' victory from the political right, left, and centre views.

#### 5.3.2.2 Annotation Method

For the frame annotation, we used a data-driven, bottom-up approach, i.e., as a group, we read each article sentence by sentence, identifying the possible frames there, observing interesting cases, and outlining the primary findings. Using the argument frame typologies proposed in the previous work (cf. Section 5.3.1), we discussed each sentence and assigned it with a respective frame type.

#### 5.3.2.3 Frame Typology

Here, we describe the frame categories we found in the articles:

- *Topical frames* which address the topic of a discussion such as "economy" and "health".
- *Value frames* which reflect personal values and beliefs such as "fairness" and "quality of life".
- *Style-based frames* which is demonstrated by the stylistic means of presenting the discussed topic. For example, this can be based on vocabulary selection in sense of using certain words and terminology, metaphors, or specific types of modalities to leverage a particular message.
- *Sentiment-based frames* relates to the choice of vocabulary that encodes a certain sentiment about a target entity or topic. Though this type can be categorised as a sub-type of *style-based frame*, we decide to consider it as a stand-alone type due to its prominence in the annotated texts. We also propose to distinguish two kinds of sentiment frames: (1) explicit sentiment that is explicitly illustrated by a chosen vocabulary (e.g., positive cheering words) and (2) implicit sentiment that centres on what "feelings" the text invokes in a reader.

### 5.3.3 Discussion

Generally speaking, developing an appropriate typology of frames requires a thorough understanding of framing and a solid theoretical ground for modelling it. Though this was the primary goal of one of the working groups in the workshop, we had limited time for collaborating and sharing the needed knowledge. Nevertheless, our pilot study illustrates several observations:

- Fame annotation is, in most cases, quite difficult and time-consuming. We have noticed that expert annotators are necessary for frame annotation, at least in the earlier stages of annotation.
- Some articles are substantially easier for identifying their frames. We assume that making the frames explicit and easy to grasp vs. making them implicit and subtle may be a strategy of the authors to convey the main message (e.g., persuade the audience).
- The frame can be represented in various forms; for instance, by a simple key phrase (e.g., fairness and equality) or by a more complex discourse relation (e.g., the contrast relation between fairness and security).
- Some articles attempt to address different frames and focus on one for them, while some completely ignore all but one of the frames.
- Frame identification, similar to many tasks in NLP but to a greater extent, is subject to the author's intent and readers' interpretation.
- Frames can be established in diverse elements of the articles, including their headlines, images with their captions, and lead paragraphs.

#### References


1 Y. Ajjour, M. Alshomary, H. Wachsmuth, and B. Stein, "Modeling frames in argumentation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 2922–2932. [Online]. Available: https://doi.org/10.18653/v1/D19-1290

2 A. E. Boydstun, D. Card, J. H. Gross, P. Resnick, and N. A. Smith, "Tracking the development of media frames within and across policy issues," in *APSA 2014 Annual Meeting Paper*, 2014.

3 D. Card, A. E. Boydstun, J. H. Gross, P. Resnik, and N. A. Smith, "The media frames corpus: Annotations of frames across issues," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015.* The Association for Computer Linguistics, 2015, pp. 438–444. [Online]. Available: https://doi.org/10.3115/v1/p15-2072

4 P. Heinisch and P. Cimiano, "A multi-task approach to argument frame classification at variable granularity levels," *it Inf. Technol.*, vol. 63, no. 1, pp. 59–72, 2021. [Online]. Available: https://doi.org/10.1515/itit-2020-0054

## 5.4 Towards Operationalizing Frames through Axiomatization

*Liesbeth Allein (KU Leuven, BE)*
*Philipp Cimiano (Universität Bielefeld, DE)*
*Ivan Habernal (TU Darmstadt, DE)*
*Martin Potthast (Leipzig University, DE)*
*Arno Simons (DZWH – Berlin, DE)*
*Benno Stein (Bauhaus-Universität Weimar, DE)*



This section summarizes the results of a working group discussion during the Dagstuhl Seminar 22131 "Framing in Communication: from Theories to Computation." Given the large number of diverging and competing frameworks and theories used to analyze framing in the computer science, the social sciences, and the digital humanities, this working group started from first principles by attempting to capture frames through axiomatization.

### 5.4.1 Axiomatization of Frames

An axiom declares a salient property of a real-world phenomenon. A set of axioms, or axiomatic systems, inductively defines the phenomenon if it entails all basic properties of the phenomenon that are not implied by others known about it. Such a set of axioms is irreducible. It opens the door to theoretical analysis of the phenomenon, i.e., the derivation of theorems that govern it. If derived theorems can be verified by observation in the real world, this raises trust in the validity of the set of axioms. A set of axioms only serves as a theoretical model of a phenomenon, if its derivable theorems predict real-world observations with sufficient accuracy.

Axiomatization relieves us from having to define frames directly, as definitions of seemingly elusive concepts, and those of frames in particular, are notoriously subject to fierce debate. In contrast, basic observations of and about frames are much less subject to debate.

As a first step towards an axiomatization of frames, we state the following three axioms:

- **Axiom 1.** Frames exist.
- **Axiom 2.** Exposure to frames has measurable effects.
- **Axiom 3.** A frame can be defined by what belongs to the frame, or by what does not belong to the frame. We call the latter a co-frame ("frame dualism").

These axioms capture fundamental prerequisites for an operationalization of frames.

Axiom 1 postulates the existence of frames in communication. Frames have a representation both in people's minds as well as in communication media, language in particular. When a frame is adopted by people, they can do this more or less reflexively (i.e., knowing that they adopt a particular frame rather than another frame). Axiom 2 postulates that frames can have measurable effects in the real world. For instance, people may change their behavior as a result of adopting a frame. Or the presence of a frame in a given piece of writing may be noticed by them. Axiom 3 postulates that the definition of a given frame can be discerned by investigating its "boundary", i.e., a frame can be discriminated from its surroundings. For example, words can be identified that have a clear connotation with the frame, or actions of people, or depictions of situations, etc. This renders frames also distinguishable from other frames, albeit interrelations between frames are not excluded. Everything that does not belong to a given frame is collectively referred to as its co-frame, thus inducing a kind of frame dualism. When dealing with an inventory of frames (i.e., string names of frames), given one frame $X$, all other frames from the inventory combined are its co-frame. A co-frame can also be described by a ranking of frames according to saliences of the frame in question. The most salient frame $Y$ of the co-frame $\bar{X}$ of frame $X$ is its main co-frame. We do not claim this set of axioms to be complete, i.e., there may be more axioms required to derive all properties of frames that have been previously observed.

From Axioms 1 and 2, it follows that the knowing or unknowing adoption of frames, and their possible measurable effects on people's behavior induces what we call camps: We define a camp by those people who share a common understanding of / recognition of / reaction to exposure to frame $X$.

As a further consequence of Axiom 2, the measurability of a frame also opens the door to its quantification. In this regard, we hypothesized that the presence of a frame may vary in terms of how well it can be recognized, called "strength" in our discussions at the time. In hindsight, a better choice of terms would have been "perceptibility", since this term comes with less ambiguous connotations.

### 5.4.2 Small Empirical Study

We conducted a brief framing perceptibility user study to support Axiom 2. We extracted six tweets about the ongoing Russian invasion of Ukraine that are said to evoke the genocide frame. Given a pair of tweets, the 21 participants of the seminar present at the time were then asked to indicate ad hoc and independent of each other in which of the two the frame of genocide is more perceptible. In our words then, which tweet contains a stronger genocide frame. We asked for participants' opinions about three pairs:

- Tweet A1 (weaker). Ex. 2:
  `A genocide that didn't happen; nuclear ambitions that Ukraine doesn't have and a threat to Russia that does not exist. I have yet hear a single justification for the murderous invasion of this country that even remotely bears scrutiny. #Ukraine`
- Tweet B1 (stronger). Ex. 8:
  `How are Western leaders sleeping during this genocide? It was posted by a woman who recorded herself right after the attack #Ukraine`
- Tweet A2 (weaker). Ex. Y:
  `Putin claims he is attacking to eliminate #Ukraine's "Nazi" government... headed by a Jewish president! Screw sanctions, Putin only cares about the price of oil. If Biden would end the insane embargo of Venezuela, oil prices would collapse and so would Putin's killing spree.`
- Tweet B2 (stronger). Ex. X:
  `#Putin is committing mass murder in #Ukraine. Why are we not doing everything in our power to stop him?`
- Tweet A3 (stronger). Ex. Z:
  `US taxpayer $$$ will fund mass murder and ethnic cleansing in my country, Ukraine.`
- Tweet B3 (weaker). Ex. 3:
  `Outrageous hypocrisy! This is the military who have committed human rights atrocities & genocide for decades in the name of "Burma's sovereignty". Both regimes must be held accountable for all serious human rights violations. #Ukraine`

Our basic operationalization of frame perceptibility (strength) to arrive at ground truth labels was this: We define the strength of a frame as the number of references to the frame. Applied to the tweets, this meant we counted the number of term occurrences which either directly refer to genocide, or have a connotation with the frame (highlighted bold), in context of what the tweet was intending to say. Tweet B3 referred to another genocide, not the one in Ukraine. The voting was as follows:

```
A1:  6 vs. B1: 15
A2:  6 vs. B2: 14 vs. Tie: 1
A3: 14 vs. B3:  6 vs. Tie: 1
```

This distribution of votes results in a Krippendorff's Alpha of -0.0349, which indicates random inter-annotator agreement and negative results for our ad hoc experiment, despite the seeming tendency of the group towards the true answer. So, while the majority decision would have been correct in all three cases, no individual annotator performed consistently well. Two comments were given by annotators: "Framing does not have ‹strength›" and "‹Stronger›is the wrong conceptualization", prompting a discussion and our change of terminology suggested above.

Axioms 2 and 3 imply that frames are discernible entities, an important prerequisite for any kind of operationalization of frames or framing. A frame provides a structure for perceiving and interpreting phenomena in a particular way. Elements of such structures can include scenes – which are themselves structures containing actors and things, and relations between actors/things – and answers to questions such as:
- What is going on?
- What is at stake?
- Who are the important actors? What are their roles? How do they relate to each other?
- What is the problem?
- What are possible solutions? What are criteria for ranking solutions?
- What can be expected to happen next?

In this regard, another more intricate operationalization of frame "perceptibility" (formerly "strength") that we conceived of was that the perceptibility of a frame is reciprocal to the number of answers it gives to the aforementioned questions.

Two frames can be compared not only in relation to the number of answers they provide, but more generally in relation to all aspects of their structure. Given frames f1 and f2, we can ask:
- Does f1 answer (some of) the same questions as f2?
- Does f1 mention (some of) the same actors? If yes, are these actors given the same or different (complementary or opposite) roles?
- Does f1 posit (some of) the same problems as f2? If yes, does f1 posit the same or different solutions to these problems? If not, are the problems posed by f1 and f2 complementary or mutually exclusive or in opposition to each other, etc.?

Comparisons of this kind can be used to provide assessments of degrees of overlap, complementarity, or mutual exclusivity of two frames.

### 5.4.3 Operationalizing Frames

The suggested three axioms above can serve as base for a number of tasks, ranging from basic tasks where two tasks are compared to more complex tasks and applications. Computational systems performing those tasks could assist stakeholders from various fields (e.g., journalists, politicians, speech writers, marketeers, educators).

Task 1 is to measure the effect of framing with all other variables fixed, i.e. we assume two texts with the same frame. The goal is to judge, first manually, then by computer models, which text is stronger. We leave the exact definition of "stronger" open for now.

In Task 2, the goal is to identify whether two texts have the same frame. Our perspective differs from the existing approaches to frame identification such that we do not rely on an existing set of predefined frame types, rather the task is simplified to a binary decision comparing two instances.

The next task, however, makes further assumptions and thus does rely on the availability of frame types. Task 3 is therefore text labeling, i.e. given a single text, does it have frame $X$? For instance, does this text describe a war as a genocide?

While the previous tasks were inherently classification-oriented, Task 4 is a text generation one. We formulate it as a text rephrasing task. Namely, given a text, we want to reframe it using a frame $X$. The open question here remains, as in the majority of text generation tasks, how to objectively evaluate the functionality.

Our list of potential tasks also includes an analysis of co-frames, where we allow for multi-label framing of texts, or an ambitious task of identifying whether a text is intentionally framed.

Given the operationalization of framing through a clear task definition, we envision the following applications. First, paraphrasing and reframing can be tailored to specific needs with respect to the audience. As with any other text generation task, a potential dual-use must be taken into account. Second, a writing assistant actively supporting framing or helping reframing a message can be beneficial in the educational context. Finally, an automated tool that highlights frames in a piece of text can help the reader to reflect on the effects of framing.

### 5.4.4 Corpus Construction for Framing Analysis

Here we outline potential strategies for compiling a corpus of annotated data. Our main presupposition is that we do not define what frames are. We propose selecting reporting on events, as framing influences how we perceive them, such as natural disasters.

Let's exemplify with a set of texts about a volcano eruption. We might observe that some of these events are labeled with extremely opposite frames, such as "disaster" and "tourist attraction". Therefore, starting with a structured database of such events uniquely identified by location and date, we might be able to sample relevant texts from social media or news. This collection would allow us to bootstrap tasks one and two, and also come up with a set of disaster-specific frame types for task three. The reframing task might be constrained in such a way that we would allow annotators only minimal lexical changes that would result in a different frame.

## Participants

- Khalid Al-Khatib
University of Groningen, NL
- Liesbeth Allein
KU Leuven, BE
- Alexander Bondarenko
Martin-Luther-Universität Halle-Wittenberg, DE
- Katarzyna Budzynska
Warsaw University of Technology, PL
- Philipp Cimiano
Universität Bielefeld, DE
- Lucie Flek
Universität Marburg, DE
- Anette Frank
Universität Heidelberg, DE
- Iryna Gurevych
TU Darmstadt, DE
- Ivan Habernal
TU Darmstadt, DE
- Annette Hautli-Janisz
Universität Passau, DE
- Zlata Kikteva
Universität Passau, DE
- Konrad Kiljan
University of Warsaw, PL
- Christopher Klamm
Universität Mannheim, DE
- Marcin Koszowy
Warsaw University of Technology, PL
- Maud Oostindie
Maastricht University, NL
- Steve Oswald
University of Fribourg, CH
- John Parkinson
Maastricht University, NL
- Martin Potthast
Universität Leipzig, DE
- Aniket Pramanick
TU Darmstadt, DE
- Chris Reed
University of Dundee, GB
- Andrea Rocci
University of Lugano, CH
- Arno Simons
DZWH – Berlin, DE
- Wassiliki Siskou
Universität Konstanz, DE
- Joanna Skolimowska
Warsaw University of Technology, PL
- Noam Slonim
IBM – Haifa, IL
- Manfred Stede
Universität Potsdam, DE
- Benno Stein
Bauhaus-Universität Weimar, DE
- Henning Wachsmuth
Universität Paderborn, DE

Report from Dagstuhl Seminar 22132

# Graph Embeddings: Theory meets Practice

**Martin Grohe*[1], Stephan Günnemann*[2], Stefanie Jegelka*[3], and Christopher Morris*[4]**

1 **RWTH Aachen University, DE.** `grohe@informatik.rwth-aachen.de`
2 **TU München, DE.** `guennemann@in.tum.de`
3 **MIT – Cambridge, US.** `stefje@csail.mit.edu`
4 **McGill University & MILA – Montreal.** `chris@christophermorris.info`

**Abstract**

Vectorial representations of graphs and relational structures, so-called graph embeddings, make it possible to apply standard tools from data mining, machine learning, and statistics to the graph domain. In particular, graph embeddings aim to capture important information about, both, the graph structure and available side information as a vector, to enable downstream tasks such as classification, regression, or clustering. Starting from the 1960s in chemoinformatics, research in various communities has resulted in a plethora of approaches, often with recurring ideas. However, most of the field advancements are driven by intuition and empiricism, often tailored to a specific application domain. Until recently, the area has received little stimulus from theoretical computer science, graph theory, and learning theory. The Dagstuhl Seminar 22132 "Graph Embeddings: Theory meets Practice", was aimed to gather leading applied and theoretical researchers in graph embeddings and adjacent areas, such as graph isomorphism, bio- and chemoinformatics, and graph theory, to stimulate an increased exchange of ideas between these communities.

**Seminar** March 27–30, 2022 – http://www.dagstuhl.de/22132
**2012 ACM Subject Classification** Mathematics of computing → Discrete mathematics; Computing methodologies → Machine learning
**Keywords and phrases** Machine Learning For Graphs, GNNs, Graph Embedding
**Digital Object Identifier** 10.4230/DagRep.12.3.141

## 1 Executive Summary

*Martin Grohe (RWTH Aachen University, DE)*
*Stephan Günnemann (TU München, DE)*
*Stefanie Jegelka (MIT – Cambridge, US)*
*Christopher Morris (McGill University & MILA – Montreal)*



Graph-structured data is ubiquitous across application domains ranging from chemo- and bioinformatics to image and social network analysis. To develop successful machine learning algorithms or apply standard data analysis tools in these domains, we need techniques that map the rich information inherent in the graph structure to a vectorial representation in a meaningful way-so-called graph embeddings. Designing such embeddings comes with unique challenges. The embedding has to account for the complex structure of (real-world) networks and additional high-dimensional continuous vectors attached to nodes and edges in a (permutation) invariant way while being scalable to massive graphs or sets of graphs. Moreover, when used in supervised machine learning, the model trained with such embeddings

* Editor / Organizer

must generalize well to new or previously unseen (graph) instances. Hence, more abstractly, designing graph embeddings results in a trade-off between expressivity, scalability, and generalization.

Starting from the 1960s in chemoinformatics, different research communities have worked in the area under various guises, often leading to recurring ideas. Moreover, triggered by the resurgence of (deep) neural networks, there is an ongoing trend in the machine learning community to design invariant/equivariant neural architectures that are capable of dealing with graph- and relational input, both (semi-)supervised and unsupervised, often denoted as graph neural networks. Although successful in practical settings, most of these developments are driven by intuition and empiricism and are geared towards specific application areas. There is no clear understanding of these approaches' limitations and their trade-offs in complexity, expressivity, and generalization. Researchers recently started to leverage connections to graph theory, group theory, logic, combinatorial algorithms, and (algorithmic) learning theory, leading to new theoretical insights and triggering new research in applications. Hence, in this seminar, we aimed to bring together leading applied and theoretical researchers in graph embeddings and adjacent areas, such as graph isomorphism, bio- and chemoinformatics, graph theory, to facilitate an increased exchange of ideas between these communities. Concretely, we aimed to understand what hinders recent theoretical developments being applied in application areas and worked towards a more practical theory. Further, we aimed at understanding the overarching challenges across applications and challenges inherent to specific areas to stimulate directions for further practical and theoretical research.

The seminar bought together 33 researchers from (applied) mathematics, specifically harmonic analysis and (algebraic) topology, (theoretical) computer science, machine learning, bioinformatics, and network science. Eighteen researchers attended remotely owing to the global COVID-19 pandemic. In total, the participants presented 18 talks on their recent progress in a better understanding of graph embeddings, focusing on supervised machine learning, particularly graph neural networks. Many talks dealt with leveraging tools from graph isomorphism testing and related areas such as finite model theory and group theory. In particular, the Weisfeiler-Leman algorithm, a popular heuristic for the graph isomorphism problem, was used to measure the expressivity of the presented algorithms and neural architectures. The consensus was that the above algorithm leads to a too coarse-grained measure of expressivity, and new notions of expressivity are needed to develop a thorough understanding. Surprisingly, only a few talks dealt with developing a better understanding of generalization, indicating that the research community still lacks an understanding. Notably, Gitta Kutyniok showed how to leverage random graph models and graphons to analyze the generalization error of graph neural networks, while Bruno Ribeiro talked about the connection between causality and out-of-distribution generalization. Further, some talks used methods from (algebraic) topology and their connection to graph theory to devise provably expressive architectures and to better understand common problems with graph neural networks, e.g., the problem of "over-smoothing" of node representations faced when considering deep architectures. Moreover, two talks covered the challenges of applying graph neural networks to biomedical data and industrial applications at Google, respectively, indicating a gap between theoretical results and practical architectures.

### Concluding Remarks

The seminar was well received, as witnessed by several positive comments from on-site participants. In general, there was an exciting atmosphere at the seminar, particularly among the large number of junior researchers attending the seminar on-site, also witnessed by many lively discussions during on-site talks. However, this was not always the case during online talks, and the active participation of online participants was relatively low. Finally, the organizers wish to express their gratitude to the Scientific Directors of Schloss Dagstuhl – Leibniz Center for Informatics for their support of the seminar.

## 2 Table of Contents

**Executive Summary**
*Martin Grohe, Stephan Günnemann, Stefanie Jegelka, and Christopher Morris* . . 141

**Overview of Talks**

Graph Neural Networks with Local Graph Parameters
*Pablo Barcelo* . . . . . . . . . . 146

Probing Graph Representations
*Aleksandar Bojchevski* . . . . . . . . . . 146

Graph Neural Networks and Graph Representation Learning Through the Lens of Curvature
*Francesco Di Giovanni* . . . . . . . . . . 147

Graph Representation Learning on Simplicial and Cellular Complexes
*Fabrizio Frasca* . . . . . . . . . . 147

Higher-order MPNNs: A Unifying Approach for Studying Expressiveness and Approximation Properties of GNNs
*Floris Geerts* . . . . . . . . . . 148

Weisfeiler and Leman Go Walking: Random Walk Kernels Revisited
*Nils Kriege* . . . . . . . . . . 148

Stability and Generalization Capabilities of Graph Neural Networks
*Gitta Kutyniok* . . . . . . . . . . 149

Equivariant Subgraph Aggregation Networks
*Haggai Maron* . . . . . . . . . . 149

Frame Averaging for Invariant and Equivariant Network Design
*Yaron Lipman* . . . . . . . . . . 150

Challenges of Applying Graph Neural Networks
*Bryan Perozzi* . . . . . . . . . . 151

Causal Graph Representation Learning
*Bruno Ribeiro* . . . . . . . . . . 151

Topology-Based Graph Learning
*Bastian Rieck* . . . . . . . . . . 152

Universal Graph Neural Networks via Random Data Augmentations Using Graph Isomorphism Tools
*Pascal Schweitzer* . . . . . . . . . . 152

Combining Representation Learning and Logical Rule Reasoning for Knowledge Graph Inference
*Yizhou Sun* . . . . . . . . . . 153

Graph Learning with 1D Convolutions on Random Walks
*Jan Tönshoff* . . . . . . . . . . 153

Graph Neural Networks are Dynamic Programmers
*Petar Velickovic* . . . . . . . . . . 154

Infusing Structure and Knowledge into Biomedical AI
*Marinka Zitnik* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 154

**Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 155

**Remote Participants** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 155

# 3 Overview of Talks

## 3.1 Graph Neural Networks with Local Graph Parameters

*Pablo Barcelo (PUC – Santiago de Chile, CL)*


**Main reference** Pablo Barceló, Floris Geerts, Juan Reutter, Maksimilian Ryschkov: "Graph Neural Networks with Local Graph Parameters", in Proc. of the Advances in Neural Information Processing Systems, Vol. 34, pp. 25280–25293, Curran Associates, Inc., 2021.
**URL** https://proceedings.neurips.cc/paper/2021/file/d4d8d1ac7e00e9105775a6b660dd3cbb-Paper.pdf

Various recent proposals increase the distinguishing power of Graph Neural Networks (GNNs) by propagating features between k-tuples of vertices. The distinguishing power of these "higher-order" GNNs is known to be bounded by the k-dimensional Weisfeiler-Leman (WL) test, yet their nonlinear memory requirements limit their applicability. Other proposals infuse GNNs with local higher-order graph structural information from the start, thereby inheriting the desirable linear memory requirement from GNNs at the cost of a one-time, possibly non-linear, preprocessing step. We propose local graph parameter enabled GNNs as a framework for studying the latter kind of approaches. We precisely characterize their distinguishing power, in terms of a variant of the WL test, and in terms of the graph structural properties that they can take into account. Local graph parameters can be added to any GNN architecture, and are cheap to compute. In terms of expressive power, our proposal lies in the middle of GNNs and their higher-order counterparts. Further, we propose several techniques to aid in choosing the right local graph parameters. Our results connect GNNs with deep results in finite model theory and finite variable logics.

## 3.2 Probing Graph Representations

*Aleksandar Bojchevski (CISPA – Saarbrücken, DE)*



Today we have a good theoretical understanding of the representational power of Graph Neural Networks (GNNs). For example, their limitations have been characterized in relation to a hierarchy of Weisfeiler-Lehman (WL) graph isomorphism tests. Consequently, there is a large body of work proposing more powerful GNNs that mitigate these limitations. We argue that these findings are only part of the story since many other factors besides the model influence learning. To complete the picture we propose a probing framework to quantify the amount of (semantically meaningful) information captured in learned graph representations. Our preliminary findings on molecular representations highlight the potential of this framework for understanding the inductive biases in GNNs and the interplay between node features and graph structure

## 3.3 Graph Neural Networks and Graph Representation Learning Through the Lens of Curvature

*Francesco Di Giovanni (Twitter – San Francisco, US)*


**Joint work of** Francesco Di Giovanni, Giulia Luise, Jake Topping, Benjamin Chamberlain, Xiaowen Dong, Michael Bronstein

Curvature is a fundamental object in the analysis of manifolds and intrinsically characterizes their geometry. It is not surprising then that synthetic notions of curvature have been introduced on graphs despite the lack of an underlying differentiable structure. In this talk, I will explore how these ideas have been recently investigated in the context of graph neural networks and graph representation learning. In the first case, curvature turns out to be the right tool to monitor the propagation of information inside message passing neural networks and allows us to properly analyse and formalize the problem of over-squashing. In the second one, we construct a family of graph embeddings into heterogeneous manifolds that are able to both match pairwise distances on the graph and the discrete graph curvature with the one on the ambient space leading to better preservation of higher order structures.

## 3.4 Graph Representation Learning on Simplicial and Cellular Complexes

*Fabrizio Frasca (Twitter – London & Imperial College London)*


**Joint work of** Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yu Guang Wang, Pietro Liò, Guido Montúfar, Michael Bronstein
**Main reference** Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yu Guang Wang, Pietro Liò, Guido Montúfar, Michael Bronstein: "Weisfeiler and Lehman Go Cellular: CW Networks", arXiv, 2021.
**URL** https://doi.org/10.48550/ARXIV.2106.12575

Graphs represent flexible and convenient mathematical abstractions for the modelling of relational systems. However, pairwise interactions may fail to capture the multi-level system of relations of many complex systems, and computational schemes embodying such paradigm are of limited expressive power. We explore topological generalisation of graphs: Simplicial and Cellular Complexes. We show they constitute natural and valid frameworks to model higher-order interactions, and how their combinatorial structure lead to the design of novel hierarchical colouring procedures extending the Weisfeiler-Leman algorithm. Graphs can be lifted to Simplicial and Cellular Complexes with appropriate transformations, allowing the application of such colouring procedures for provably more expressive representations. Finally, these procedures inspire the design of neural counterparts implementing a form of higher-order message passing. These expressive architectures overcome several limitations of standard Graph Neural Networks; we show they excel on a variety of graph learning benchmarks and obtain state-of-the-art results on various molecular datasets.

## 3.5 Higher-order MPNNs: A Unifying Approach for Studying Expressiveness and Approximation Properties of GNNs

*Floris Geerts (University of Antwerp, BE)*


**Joint work of** Floris Geerts, Juan L Reutter
**Main reference** Floris Geerts, Juan L Reutter: "Expressiveness and Approximation Properties of Graph Neural Networks", in Proc. of the International Conference on Learning Representations, 2022.
**URL** https://openreview.net/forum?id=wIzUeM3TAU

Characterizing the separation power of graph neural networks (GNNs) provides an understanding of their limitations for graph learning tasks. Results regarding separation power are, however, usually geared at specific GNN architectures, and tools for understanding arbitrary GNN architectures are generally lacking. We provide an elegant way to easily obtain bounds on the separation power of GNNs in terms of the Weisfeiler-Leman (WL) tests, which have become the yardstick to measure the separation power of GNNs. The crux is to view GNNs as expressions in a procedural tensor language describing the computations in the layers of the GNNs. Then, by a simple analysis of the obtained expressions, in terms of the number of indexes and the nesting depth of summations, bounds on the separation power in terms of the WL-tests readily follow. We use tensor language to define Higher-Order Message-Passing Neural Networks (or k-MPNNs), a natural extension of MPNNs. Furthermore, the tensor language point of view allows for the derivation of universality results for classes of GNNs in a natural way. Our approach provides a toolbox with which GNN architecture designers can analyze the separation power of their GNNs, without needing to know the intricacies of the WL-tests. We also provide insights in what is needed to boost the separation power of GNNs.

## 3.6 Weisfeiler and Leman Go Walking: Random Walk Kernels Revisited

*Nils Kriege (Universität Wien, AT)*


**Main reference** Nils M. Kriege: "Weisfeiler and Leman Go Walking: Random Walk Kernels Revisited", CoRR, Vol. abs/2205.10914, 2022.
**URL** https://doi.org/10.48550/arXiv.2205.10914

Random walk kernels have been introduced in seminal work on graph learning and were later largely superseded by kernels based on the Weisfeiler-Leman test for graph isomorphism. We give a unified view on both classes of graph kernels. We study walk-based node refinement methods and formally relate them to several widely-used techniques, including Morgan's algorithm for molecule canonization and the Weisfeiler-Leman test. We define corresponding walk-based kernels on nodes that allow fine-grained parameterized neighborhood comparison, reach Weisfeiler-Leman expressiveness, and are computed using the kernel trick. From this we show that classical random walk kernels with only minor modifications regarding definition and computation are as expressive as the widely-used Weisfeiler-Leman subtree kernel but support non-strict neighborhood comparison. We verify experimentally that walk-based kernels reach or even surpass the accuracy of Weisfeiler-Leman kernels in real-world classification tasks.

## 3.7 Stability and Generalization Capabilities of Graph Neural Networks

*Gitta Kutyniok (LMU München, DE)*


**Joint work of** Gitta Kutyniok, Holger Boche, Michael M. Bronstein, Lorenzo Bucci, Adalbert Fono, Wei Huang, Yunseok Lee, Ron Levie, Sohir Maskey

The tremendous importance of graph structured data due to recommender systems or social networks led to the introduction of graph convolutional neural networks (GCN). We ask the question to which extent GCN are able to generalize to graphs, which describe a similar phenomenon as present in the training data set. We consider different notions of similarity, using random graph models as well as graphons, and analyze both spectral GCNs [1, 2] and message passing neural networks [3]. In these settings, we will then derive comprehensive non-asymptotic bounds on the related generalization error. We will finish with a word of caution when training graph neural networks on classical digital hardware, and present fundamental limitations [4, 5].

### References

1 R. Levie, W. Huang, L. Bucci, M. M. Bronstein, and G. Kutyniok. Transferability of Spectral Graph Convolutional Neural Networks. J. Mach. Learn. Res., to appear. (arXiv:1907.12972).
2 S. Maskey, R. Levie, and G. Kutyniok. Transferability of Graph Neural Networks: an Extended Graphon Approach. (arXiv:2109.10096)
3 S. Maskey, Y. Lee, R. Levie, and G. Kutyniok. Stability and Generalization Capabilities of Message Passing Graph Neural Networks (arXiv:2202.00645)
4 H. Boche, A. Fono and G. Kutyniok. Limitations of Deep Learning for Inverse Problems on Digital Hardware (arXiv:2202.13490)
5 H. Boche, A. Fono and G. Kutyniok. Inverse Problems Are Solvable on Real Number Signal Processing Hardware (arxiv:2204.02066)

## 3.8 Equivariant Subgraph Aggregation Networks

*Haggai Maron (NVIDIA – Tel Aviv, IL)*


**Joint work of** Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M. Bronstein, Haggai Maron
**Main reference** Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M. Bronstein, Haggai Maron: "Equivariant Subgraph Aggregation Networks", in Proc. of the International Conference on Learning Representations, 2022.
**URL** https://openreview.net/forum?id=dFbKQaRk15w

Message-passing neural networks (MPNNs) are the leading architecture for deep learning on graph-structured data, in large part due to their simplicity and scalability. Unfortunately, it was shown that these architectures are limited in their expressive power. This work proposes a novel framework called Equivariant Subgraph Aggregation Networks (ESAN) to address this issue. Our main observation is that while two graphs may not be distinguishable by an MPNN, they often contain distinguishable subgraphs. Thus, we propose to represent each graph as a set of subgraphs derived by some predefined policy, and to process it using a suitable equivariant architecture. We develop novel variants of the 1-dimensional Weisfeiler-Leman (1-WL) test for graph isomorphism, and prove lower bounds on the expressiveness of ESAN in terms of these new WL variants. We further prove that our approach increases the expressive power of both MPNNs and more expressive architectures. Moreover, we provide theoretical results that describe how design choices such as the subgraph selection policy and equivariant neural architecture affect our architecture's expressive power. To deal with the increased computational cost, we propose a subgraph sampling scheme, which can be viewed as a stochastic version of our framework. A comprehensive set of experiments on real and synthetic datasets demonstrates that our framework improves the expressive power and overall performance of popular GNN architectures.

## 3.9 Frame Averaging for Invariant and Equivariant Network Design

*Yaron Lipman (Weizmann Institute – Rehovot, IL)*

**Joint work of** Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, Yaron Lipman

**Main reference** Omri Puny, Matan Atzmon, Heli Ben-Hamu, Edward J. Smith, Ishan Misra, Aditya Grover, Yaron Lipman: "Frame Averaging for Invariant and Equivariant Network Design", CoRR, Vol. abs/2110.03336, 2021.
**URL** https://arxiv.org/abs/2110.03336

Many machine learning tasks involve learning functions that are known to be invariant or equivariant to certain symmetries of the input data. However, it is often challenging to design neural network architectures that respect these symmetries while being expressive and computationally efficient. For example, Euclidean motion invariant/equivariant graph or point cloud neural networks.

In this work we introduce Frame Averaging (FA), a general purpose and systematic framework for adapting known (backbone) architectures to become invariant or equivariant to new symmetry types. Our framework builds on the well known group averaging operator that guarantees invariance or equivariance but is intractable. In contrast, we observe that for many important classes of symmetries, this operator can be replaced with an averaging operator over a small subset of the group elements, called a frame. We show that averaging over a frame guarantees exact invariance or equivariance while often being much simpler to compute than averaging over the entire group. Furthermore, we prove that FA-based models have maximal expressive power in a broad setting and in general preserve the expressive power of their backbone architectures. Using frame averaging, we propose a new class of universal Graph Neural Networks (GNNs), universal Euclidean motion invariant point cloud networks, and Euclidean motion invariant Message Passing (MP) GNNs. We demonstrate the practical effectiveness of FA on several applications including point cloud normal estimation, beyond 2-WL graph separation, and $n$-body dynamics prediction, achieving state-of-the-art results in all of these benchmarks.

## 3.10 Challenges of Applying Graph Neural Networks

*Bryan Perozzi (Google – New York, US)*


**Joint work of** Bryan Perozzi, John Palowitch, Anton Tsitsulin, Brandon Mayer, Qi Zhu, Natalia Ponomareva, Jonathan Halcrow, Sam Ruth, Alexandru Mosoi, Jiawei Han
**Main reference** Jonathan Halcrow, Alexandru Mosoi, Sam Ruth, Bryan Perozzi: "Grale: Designing Networks for Graph Learning", in Proc. of the KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, pp. 2523–2532, ACM, 2020.
**URL** https://doi.org/10.1145/3394486.3403302

Graph Neural Networks are a tantalizing way of modeling data which doesn't have a fixed structure. However, getting them to work as expected has had some twists and turns over the years. In this talk, I discuss three efforts from our group on important (and understudied) problems applying GNNs to real data including graph construction, model benchmarking, and model robustness:

Grale, is a scalable method we have developed to address the problem of graph design for graphs with billions of nodes. Grale operates by fusing together different measures of (potentially weak) similarity to create a graph which exhibits high task-specific homophily between its nodes. Grale is designed for running on large datasets. We have deployed Grale in more than 20 different industrial settings at Google, including datasets which have tens of billions of nodes, and hundreds of trillions of potential edges to score.

GraphWorld is a novel methodology and system for benchmarking GNN models on an arbitrarily-large population of synthetic graphs for any conceivable GNN task. GraphWorld allows a user to efficiently generate a world with millions of statistically diverse datasets. It is accessible, scalable, and easy to use. GraphWorld can be run on a single machine without specialized hardware, or it can be easily scaled up to run on arbitrary clusters or cloud frameworks. Using GraphWorld, a user has fine-grained control over graph generator parameters, and can benchmark arbitrary GNN models with built-in hyperparameter tuning

Shift-Robust GNN (SR-GNN) is designed to account for distributional differences between biased training data and a graph's true inference distribution. SR-GNN adapts GNN models to the presence of distributional shift between the nodes labeled for training and the rest of the dataset. We illustrate the effectiveness of SR-GNN in a variety of experiments with biased training datasets on common GNN benchmark datasets for semi-supervised learning, where we see that SRGNN outperforms other GNN baselines in accuracy, addressing at least ~40% of the negative effects introduced by biased training data.

## 3.11 Causal Graph Representation Learning

*Bruno Ribeiro (Purdue University – West Lafayette, US)*


**Joint work of** Bruno Ribeiro, Beatrice Bevilacqua, Yangze Zhou, S Chandra Mouli

In this talk I discussed the challenges and opportunities in building graph representations for causal tasks (learning and prediction). We started with the question "Why is causality relevant for graph machine learning?", expanding it into three threads: (a) Some graph tasks are causal, such as link prediction for recommender systems; (b) Extrapolation tasks in deep learning better defined through causality, since convex hull and other geometric definitions in high dimensions tend to be meaningless for machine learning; (c) Out-of-distribution tasks are a mix of associational and counterfactual tasks (as the work of Bevilacqua et al. 2021 and Mouli et al. 2021 show). For out-of-distribution tasks we reviewed the concept of counterfactual invariant (graph) representations (Bevilacqua et al. 2021). Explaining why data augmentations for graphs are difficult to properly implement in practice (e.g., what it would look like if graph were larger without changing class label?). The talk ended stating that counterfactual-invariant representations are task-dependent and that, unlike associational graph tasks, there are provably no universal approximators for causal tasks.

## 3.12 Topology-Based Graph Learning

*Bastian Rieck (Helmholtz Zentrum München, DE)*


**Joint work of** Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, Karsten Borgwardt
**Main reference** Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, Karsten Borgwardt: "Topological Graph Neural Networks", in Proc. of the International Conference on Learning Representations, 2022.
**URL** https://openreview.net/forum?id=oxxUMeFwEHd

Topological data analysis is starting to establish itself as a powerful and effective framework in machine learning, supporting the analysis of neural networks, but also driving the development of novel algorithms that incorporate topological characteristics. As a problem class, graph representation learning is of particular interest here, since graphs are inherently amenable to a topological description in terms of their connected components and cycles. This talk will provide an overview of how to address graph learning tasks using machine learning techniques, with a specific focus on how to make such techniques 'topology-aware.' We will discuss how to learn filtrations for graphs and how to incorporate topological information into modern graph neural networks, resulting in provably more expressive algorithms. This talk aims to be accessible to an audience of graph learning enthusiasts; prior knowledge of topological data analysis is helpful but not required.

## 3.13 Universal Graph Neural Networks via Random Data Augmentations Using Graph Isomorphism Tools

*Pascal Schweitzer (TU Darmstadt, DE)*


**Joint work of** Pascal Schweitzer, Markus Anders, Billy Joe Franks, Marius Kloft

Message-passing neural networks have provable limitations. Random data augmentations can be used to overcome these, resulting in provably universal graph neural networks. I will describe a solver from the realm of practical graph isomorphism testing that is based on so-called individualization-refinement techniques and uses random sampling. I will then describe how it can be employed to obtain efficient, scalable, universal graph neural networks.

## 3.14 Combining Representation Learning and Logical Rule Reasoning for Knowledge Graph Inference

*Yizhou Sun (UCLA, US)*


**Joint work of** Xuelu Chen, Ziniu Hu, Kewei Cheng, Ziqing Yang, Ming Zhang
**Main reference** Xuelu Chen, Ziniu Hu, Yizhou Sun: "Fuzzy Logic Based Logical Query Answering on Knowledge Graphs", in Proc. of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 – March 1, 2022, pp. 3939–3948, AAAI Press, 2022.
**URL** https://ojs.aaai.org/index.php/AAAI/article/view/20310
**Main reference** Kewei Cheng, Ziqing Yang, Ming Zhang, Yizhou Sun: "UniKER: A Unified Framework for Combining Embedding and Definite Horn Rule Reasoning for Knowledge Graph Inference", in Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pp. 9753–9771, Association for Computational Linguistics, 2021.
**URL** https://doi.org/10.18653/v1/2021.emnlp-main.769

Knowledge graph inference has been studied extensively due to its wide applications. It has been addressed by two lines of research, i.e., the more traditional logical rule reasoning and the more recent knowledge graph embedding (KGE). In this talk, we will introduce two recent developments in our group to combine these two worlds. First, we propose to leverage logical rules to bring in high-order dependency among entities and relations for KGE. By limiting the logical rules to be the definite Horn clauses, we are able to fully exploit the knowledge in logical rules and enable the mutual enhancement of logical rule-based reasoning and KGE in an extremely efficient way. Second, we propose to handle logical queries by representing fuzzy sets as specially designed vectors and retrieving answers via dense vector computation. In particular, we provide embedding-based logical operators that strictly follow the axioms required in fuzzy logic, which can be trained by self-supervised knowledge completion tasks. With additional query-answer pairs, the performance can be further enhanced. With these evidence, we believe combining logic with representation learning provides a promising direction for knowledge reasoning.

## 3.15 Graph Learning with 1D Convolutions on Random Walks

*Jan Tönshoff (RWTH Aachen, DE)*


**Joint work of** Jan Tönshoff, Martin Ritzert, Hinrikus Wolf, Martin Grohe
**Main reference** Jan Tönshoff, Martin Ritzert, Hinrikus Wolf, Martin Grohe: "Graph Learning with 1D Convolutions on Random Walks", CoRR, Vol. abs/2102.08786, 2021.
**URL** https://arxiv.org/abs/2102.08786

We propose CRaWl (CNNs for Random Walks), a novel neural network architecture for graph learning. It is based on processing sequences of small subgraphs induced by random walks with standard 1D CNNs. Thus, CRaWl is fundamentally different from typical message passing graph neural network architectures. It is inspired by techniques counting small subgraphs, such as the graphlet kernel and motif counting, and combines them with random walk based techniques in a highly efficient and scalable neural architecture. We demonstrate empirically that CRaWl matches or outperforms state-of-the-art GNN architectures across a multitude of benchmark datasets for graph learning.

## 3.16 Graph Neural Networks are Dynamic Programmers

*Petar Velickovic (DeepMind – London, GB)*


**Joint work of** Andrew Dudzik, Petar Veličković
**Main reference** Andrew Dudzik, Petar Veličković: "Graph Neural Networks are Dynamic Programmers", CoRR, Vol. abs/2203.15544, 2022.
**URL** https://doi.org/10.48550/arXiv.2203.15544

Recent advances in neural algorithmic reasoning with graph neural networks (GNNs) are propped up by the notion of algorithmic alignment. Broadly, a neural network will be better at learning to execute a reasoning task (in terms of sample complexity) if its individual components align well with the target algorithm. Specifically, GNNs are claimed to align with dynamic programming (DP), a general problem-solving strategy which expresses many polynomial-time algorithms. However, has this alignment truly been demonstrated and theoretically quantified? Here we show, using methods from category theory and abstract algebra, that there exists an intricate connection between GNNs and DP, going well beyond the initial observations over individual algorithms such as Bellman-Ford. Exposing this connection, we easily verify several prior findings in the literature, and hope it will serve as a foundation for building stronger algorithmically aligned GNNs.

## 3.17 Infusing Structure and Knowledge into Biomedical AI

*Marinka Zitnik (Harvard University – Boston, US)*


**Joint work of** Xiang Zhang, Kexin Huang, Michelle Li, Qianwen Wang, Joshua Pan, Theodoros Tsiligkaridis, Emily Alsentzer, Sam Finlayson, Matthew McDermott, Joe Loscalzo, Jure Leskovec, Laszlo Barabasi, Marinka Zitnik

Artificial intelligence has enabled scientific breakthroughs in diverse areas of biology and medicine. However, biomedical data present unique challenges, including limited annotations for supervised learning, the need to generalize to new scenarios not seen during training, and the need for trustworthy representations that lend themselves to actionable hypotheses in the laboratory. This talk describes our efforts to address these challenges by infusing structure and knowledge into biomedical AI. First, I outline subgraph neural networks that can disentangle distinct aspects of subgraph structure. I will then present a general-purpose approach for few-shot learning on graphs. At the core is the notion of local subgraphs that transfer knowledge from one task to another, even when only a handful of labeled examples are available. This principle is theoretically justified as we show that the evidence for predictions can be found in subgraphs surrounding the targets. Finally, to illustrate the benefits of modeling structure in non-graph datasets, I will introduce Raindrop, a graph neural network that embeds complex time series while also learning the dynamics of sensors purely from observational data. This research creates new avenues for accelerating drug discovery, fusing biomedical knowledge and patient data, and giving the right patient the right treatment at the right time to have effects that are consistent from person to person and with results in the laboratory.

## Participants

- Francesco Di Giovanni
Twitter – San Francisco, US
- Federico Errica
NEC Laboratories Europe – Heidelberg, DE
- Fabrizio Frasca
Twitter – London & Imperial College London, UK
- Floris Geerts
University of Antwerp, BE
- Martin Grohe
RWTH Aachen University, DE
- Stephan Günnemann
TU München, DE
- Nils Kriege
Universität Wien, AT
- Gitta Kutyniok
LMU München, DE
- Haggai Maron
NVIDIA – Tel Aviv, IL
- Christopher Morris
McGill University & MILA – Montreal
- Gaurav Rattan
RWTH Aachen, DE
- Bruno Ribeiro
Purdue University – West Lafayette, US
- Bastian Rieck
Helmholtz Zentrum München, DE
- Pascal Schweitzer
TU Darmstadt, DE
- Jan Tönshoff
RWTH Aachen, DE

## Remote Participants

- Pablo Barcelo
PUC – Santiago de Chile, CL
- Aleksandar Bojchevski
CISPA – Saarbrücken, DE
- Joan Bruna Estrach
New York University, US
- Tina Eliassi-Rad
Northeastern University – Boston, US
- Matthias Fey
TU Dortmund, DE
- Barbara Hammer
Universität Bielefeld, DE
- Stefanie Jegelka
MIT – Cambridge, US
- Elias Khalil
University of Toronto, CA
- Benny Kimelfeld
Technion – Haifa, IL
- Yaron Lipman
Weizmann Institute – Rehovot, IL
- Andreas Loukas
Prescient Design – Schlieren, CH
- Bryan Perozzi
Google – New York, US
- Siamak Ravanbakhsh
McGill University – Montréal, CA
- Joshua Robinson
MIT – Cambridge, US
- Yizhou Sun
UCLA, US
- Petar Velickovic
DeepMind – London, GB
- Ulrike von Luxburg
Universität Tübingen, DE
- Marinka Zitnik
Harvard University – Boston, US