DAGSTUHL
REPORTS

Volume 14, Issue 8, August 2024

 ISSN 2192-5283

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
March, 2025

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.14.8.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 24341

Proof Representations: From Theory to Applications

Anupam Das*[1], **Elaine Pimentel***[2], **Lutz Straßburger***[3], **and Robin Martinot**†[4]

1 **University of Birmingham, GB.** a.das@bham.ac.uk
2 **University College London, GB.** e.pimentel@ucl.ac.uk
3 **INRIA Saclay – Île-de-France, FR.** lutz@lix.polytechnique.fr
4 **Utrecht University, NL.** r.a.martinot@uu.nl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 24341 "Proof Representations: From Theory to Applications". Proof theory is the study of formal proofs as mathematical objects in their own right. The subject has enjoyed continued attention among computer scientists in particular due to its significance for formalization, metalogic, and automation. In recent decades there has been a surge of interest on the representations of formal proofs themselves. The outcomes of these investigations have been remarkable, in particular extending the scope of structural proof theory to novel and richer settings:

Richer line structures (e.g. hypersequents, nested sequents, labelled sequents) have resulted in a uniform treatment of standard modal logics, streamlining their metatheory and providing new tools for metalogical problems. Richer proof structures (e.g., cyclic proofs, annotated systems, infinitely branching proofs) have significantly advanced our understanding of fixed points and (co)induction. Indeed, we are now seeing many of these previously disjoint techniques being combined to push the boundaries of proof theoretic approaches to computational logic. Graphical proof representations (e.g., proof nets, atomic flows, combinatorial proofs) originating from "linear" logics, now not only comprise a well-behaved computational model for resource-sensitive reasoning, but also provide an impressively uniform treatment for logics across the board. In fact, we are now seeing many of these previously disjoint techniques being combined to push the boundaries of proof theoretic approaches to computational logic, which has produced deep and fruitful cross-fertilizations between programming languages and proof theory. Arguably, the most well-known is the Curry-Howard correspondence ("propositions-as-types") where (functional) programs correspond to formal proofs and their execution to normalization. A complementary tradition, proof-search-as-computation ("propositions-as-processes"), instead interprets (logic) programs to formulas and their execution to proof search.

The goal of this Dagstuhl Seminar was twofold. First and foremost, we aimed to bring together theorists and practitioners exploiting proof representations to identify new directions of application and, simultaneously, distill new theoretical directions from problems "in the wild". At the same time, this seminar was intended to expose the interface between the proof-normalization and proof-search traditions by probing proof representations from both directions.

Seminar August 18–23, 2024 – https://www.dagstuhl.de/24341
2012 ACM Subject Classification Theory of computation → Logic; Theory of computation → Proof theory; Theory of computation → Automated reasoning
Keywords and phrases proof theory, proof calculi, computational interpretations, proof semantics, dynamic operators
Digital Object Identifier 10.4230/DagRep.14.8.1

* Editor / Organizer
† Editorial Assistant / Collector

Proof Representations: From Theory to Applications, *Dagstuhl Reports*, Vol. 14, Issue 8, pp. 1–23
Editors: Anupam Das, Elaine Pimentel, and Lutz Straßburger
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Anumpam Das (University of Birmingham, GB, a.das@bham.ac.uk)
Elaine Pimentel (University College London, GB, e.pimentel@ucl.ac.uk)
Lutz Straßburger (INRIA Saclay – Île-de-France, FR, lutz@lix.polytechnique.fr)

Dagstuhl Seminar 24341 was organized in response to growing interest in the representation of formal proofs. Its primary aim was to bring together theorists and practitioners who are exploring various proof representations, with the goal of identifying new applications while simultaneously creating new theoretical directions. A key focus of the seminar was to explore the interface between proof normalization and proof search traditions, by examining proof representations from both perspectives.

The seminar focused on the relation between various new developments in the field, including: the more philosophical direction of proof-theoretic semantics; the upcoming unifying research program of universal proof theory; and the study of non-standard proof formats, such as non-well–founded proofs. Individual talks covered a broad array of topics relevant to these developments, such as the particulars of certain variants on standard proof calculi (calculi including witness operators, calculi for inconsistent logics and logics with numerical quantification, calculi for second-order logic, and "exotic" proof calculi, e.g. labelled calculi for modal and intuitionistic logics), but also various connections of proof theory to computational applications (verification logic, propositional dynamic logic, tableaux using SAT solvers, and interactive theorem provers). To do justice to these complex topics, researchers from proof theory, computational logic, and philosophical logic were invited to collaborate and provide insights. The seminar also identified several research gaps across these areas. One important takeaway was the importance of utilizing different representations of proof systems to address emerging open questions in the field, and to prioritize the need for unification.

The seminar itself was structured to encourage extensive interaction among participants, both formally and informally. Participants were given considerable freedom in preparing their contributions and during the seminar itself, including the option to give talks in a variety of formats. Several in-depth special sessions were organized to focus on the most important developments in the field, and longer individual talks by experts covered specific topics in greater detail. Evening sessions, dubbed "beer talks", provided a relaxed environment for casual discussions, allowing participants to explore topics of shared interest and build connections across different areas of research.

The discussions from the seminar reflect the key interests of the community and provide valuable topics for ongoing research. At the end of the seminar, participants agreed to stay in contact to continue their discussions and foster new collaborations. Already various initiatives are being taken. For instance, Fernando Raymundo Velazquez Quesada, Carlos Olarte, and Elaine Pimentel began a promising collaboration on Epistemic Propositional Dynamic Logic following discussions at Dagstuhl. They recently held an in-person meeting in Bergen, Norway, and plan to apply for a European grant in the near future. Additionally, collaborations have among others been started by Lutz Straßburger and Revantha Ramanayake on modal logic, and by Lutz Straßburger and Matteo Acclavio on linear logic. We look forward to seeing how the results of these efforts will further shape the field.

2 Table of Contents

3 Overview of Talks

3.1 Even more exotic proof systems

Matteo Acclavio (University of Southern Denmark – Odense, DK)

In this talk, I presented several proof systems that have been underrepresented throughout the week. These systems are categorized into three groups, each with its own strengths and weaknesses: (1) proofs as sequences of rules (e.g., sequent calculi, natural deduction, and deep inference); (2) proofs as argumentation (e.g., dialogical games, Hyland-Ong games); (3) proofs as sound relations between elements of a theorem (e.g., proof nets, combinatorial proofs, and connection method proofs).

3.2 Fantastic connectives and where to find them

Matteo Acclavio (University of Southern Denmark – Odense, DK)

In this talk, I discuss the notion of (multiplicative) logical connectives. I address the limitations of connectives defined by rules, such as synthetic connectives, and generalized connectives from the early works in linear logic (defined as sets of partitions). I then briefly explain how defining connectives as prime graphs enables the development of logics and proof systems with a more robust proof theory.

References

1 Matteo Acclavio and Roberto Maieli. *Generalized Connectives for Multiplicative Linear Logic*. https://doi.org/10.4230/LIPIcs.CSL.2020.6
2 Matteo Acclavio, Ross Horne and Lutz Strassburger. *An Analytic Propositional Proof System on Graphs*. https://doi.org/10.46298/lmcs-18(4:1)2022
3 Matteo Acclavio and Roberto Maieli. *Logic Programming with Multiplicative Structures*. https://doi.org/10.4204/EPTCS.408.3
4 Matteo Acclavio. *Sequent Systems on Undirected Graphs*. https://doi.org/10.1007/978-3-031-63501-4_12

3.3 Demystifying mu

Bahareh Afshari (University of Gothenburg, SE)

Joint work of Bahareh Afshari, Graham E. Leigh, Guillermo Menéndez Turata
Main reference Bahareh Afshari, Graham E. Leigh, Guillermo Menéndez Turata: "Demystifying μ", CoRR, Vol. abs/2401.01096, 2024.
URL https://doi.org/10.48550/ARXIV.2401.01096

Formal proofs are an important tool in the mathematical study of computational systems, such as certifying the correctness of an abstract model or verifying that an implementation adheres to a specification. The certification and verification role of proofs boils down to questions of proof existence and proof synthesis: Does a given formula (sequent, judgement, etc.) admit a proof? Can a proof be generated for each provable formula (sequent, etc.)? Both questions rest heavily on a third, often unspoken: what constitutes a proof?

A proof is normally understood as a finite tree with vertices labelled from a class of deduction elements (formulas, sequents, judgements, etc.) where each vertex together with its children match the conclusion and premises of one of a fixed collection of inferences, the decision of which should be complexity-theoretically simple, at worst polynomial-time decidable. This definition is computationally sufficient for many logics such as propositional logic, modal logics, even predicate logic. In each case, soundness and completeness theorems for the corresponding notion of proof confirm that proof existence agrees with semantic validity, and normal form theorems – notably admissibility of cut – give rise to proof synthesis.

Logics and theories incorporating inductive and co-inductive concepts strain the traditional notion of proof, however. Proof existence relies on more complex completeness theorems which coax the infinitary behaviour of the inductive and co-inductive constructions into finitary induction principles. Synthesis likewise suffers, becoming a task in discerning induction invariants from mere provability or elaborating the proof calculus to recover desirable normal forms. The latter is the typical approach of sequent calculi for modal logics where inductive properties – transitivity, factivity, well-foundedness, etc. – are internalised in the inferences rules.

Illfounded proofs offer to alleviate the proof-theorist's burden not by complicating the notion of proofs but by relaxing it. Proofs are no longer constrained to finite trees; they can be infinite trees, or even graphs, at the cost of infinite branches fulfilling some pre-determined correctness condition. Illfounded proofs provide an alternative to the traditional notion of proof that treats logics and theories of inductive concepts as manifestations of a general infinitary framework rather than diverging extensions of finitary logic. With a change in perspective comes the need to revisit the encompassing theory. Indeed, the inductive construction of formal proof plays a non-trivial role in most – if not all – fundamental results of well-founded proof theory: proof transformations, cut elimination, computational interpretations, syntactic approaches to interpolation and so on.

The talk presented an introduction to the theory and application of illfounded and cyclic proof systems. We focused on two aspects, interpolation and completeness, and saw how results concerning illfounded proofs can be reflected to finitary, inductive, proofs.

3.4 Abella: An Overview

Kaustuv Chaudhuri (INRIA Saclay – Île-de-France, FR)

Joint work of Kaustuv Chaudhuri, David Baelde, Andrew Gacek, Dale Miller, Gopalan Nadathur, Alwen Tiu, Yuting Wang
Main reference David Baelde, Kaustuv Chaudhuri, Andrew Gacek, Dale Miller, Gopalan Nadathur, Alwen Tiu, Yuting Wang: "Abella: A System for Reasoning about Relational Specifications", J. Formaliz. Reason., Vol. 7(2), pp. 1–89, 2014.
URL https://doi.org/10.6092/ISSN.1972-5787/4650

Abella is an interactive theorem prover for reasoning about higher-order relational specifications. It is particularly well suited for formalizing the meta-theory of deductive formalisms, including proof systems, type systems, operational semantics, process calculi, etc. One of its key features is the ability to define the inductive structure of data with binding constructs such as quantifiers from logic, abstraction from programming languages, and restriction from process calculi.

This talk gives an overview of the system and presents a main example that can be seen as a challenge problem for formalized meta-theory: showing the equivalence of the higher-order and De Bruijn indexed representations of λ-terms.

3.5 Direct Manipulation for Interactive Theorem Proving

Kaustuv Chaudhuri (INRIA Saclay – Île-de-France, FR)

Main reference Kaustuv Chaudhuri: "Subformula Linking for Intuitionistic Logic with Application to Type Theory", in Proc. of the Automated Deduction - CADE 28 - 28th International Conference on Automated Deduction, Virtual Event, July 12-15, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12699, pp. 200–216, Springer, 2021.
URL https://doi.org/10.1007/978-3-030-79876-5_12

Interactive theorem provers commonly have formal proof languages for expressing formal proofs. Such languages are rarely portable across different systems. I propose an alternative interactive proving interface where textual proof languages have a negligible role. Instead, proofs are built by manipulating the theorem using interaction devices such as mice and touch screens, and interaction mechanisms such as clicking, dragging-and-dropping, etc. The proofs can then be extracted in a variety of existing textual languages. This talk covers intuitionistic first-order logic and discusses some open challenges for this technique, particularly for dependent type theory.

A web-based prototype implementation can be found at: `https://chaudhuri.info/research/profint/index-dagstuhl-2024.html`.

3.6 Exotic proof systems

Marianna Girlando (University of Amsterdam, NL)

In this tutorial we explore the proof theory of modal logic. It is well-known that Gentzen-style sequent calculi fail to meet basic requirements in the case of modal logic: for instance, no cut-free sequent calculus is known for the modal logic S5. In order to overcome these difficulties, several extensions of Gentzen's formalism, which we refer to as "exotic" proof systems, have been studied in the literature. These can be approximately grouped into two categories: labelled calculi, which extend the language of the calculus with explicit semantic information, and structured calculi, among which nested sequents, which instead enrich the structure of sequents. In this tutorial, we will cover the basics of labelled calculi and nested calculi for the S5-cube of modal logics.

3.7 CEGAR-Tableaux: improved modal satisfiability via modal clause-learning and SAT

Rajeev P. Gore (Turner, AU)

Joint work of Rajeev Gore, Cormac Kikkert

We present CEGAR-Tableaux, a tableaux-like method for propositional modal logics utilising SAT-solvers, modal clause-learning and multiple optimisations from modal and description logic tableaux calculi. We use the standard Counter-example Guided Abstract Refinement (CEGAR) strategy for SAT-solvers to mimic a tableau-like search strategy that explores a rooted tree-model with the classical propositional logic part of each Kripke world evaluated using a SAT-solver. Unlike modal SAT-solvers and modal resolution methods, we do not explicitly represent the accessibility relation but track it implicitly via recursion. By using "satisfiability under unit assumptions", we can iterate rather than "backtrack" over the satisfiable diamonds at the same modal level (context) of the tree model with one SAT-solver. By keeping modal contexts separate from one another, we add further refinements for reflexivity and transitivity which manipulate modal contexts once only. Our solver CEGARBox is, overall, the best for modal logics K, KT and S4 over the standard benchmarks, sometimes by orders of magnitude.

3.8 Proof systems for term-forming operators

Andrzej Indrzejczak (University of Lodz, PL)

Main reference Andrzej Indrzejczak: "Towards Proof-Theoretic Formulation of the General Theory of Term-Forming Operators", in Proc. of the Automated Reasoning with Analytic Tableaux and Related Methods - 32nd International Conference, TABLEAUX 2023, Prague, Czech Republic, September 18-21, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14278, pp. 131–149, Springer, 2023.
URL https://doi.org/10.1007/978-3-031-43513-3_8

Complex descriptive terms are commonly used in natural languages but its role in conveying information is rather neglected in formal languages. This short talk presents the possible way of constructing well-behaved proof systems for operators enabling the expression of complex terms (like definite descriptions, set-abstracts) in proof systems, like sequent calculi, natural deduction, tableaux systems etc.[1]

[1] This work is sponsored by the ERC advanced grant ERC-2021-ADG, ExtenDD 101054714.

3.9 Epsilon Calculus and LK

Anela Lolic (TU Wien, AT)

Joint work of Anela Lolic, Matthias Baaz, Alexander Leitsch
Main reference Kaustuv Chaudhuri: "Subformula Linking for Intuitionistic Logic with Application to Type Theory", in Proc. of the Automated Deduction - CADE 28 - 28th International Conference on Automated Deduction, Virtual Event, July 12-15, 2021, Proceedings, Lecture Notes in Computer Science, Vol. 12699, pp. 200–216, Springer, 2021.
URL https://doi.org/10.1007/978-3-030-79876-5_12

The epsilon calculus, which is the oldest formalism in proof theory, provides important algorithms for proof theoretic properties through its theorems. For example, Herbrand disjunctions of existential formulas can be computed via the extended first epsilon theorem. The proof of the second epsilon theorem provides a direct approach for the elimination of Skolem functions from Herbrand disjunctions of existential formulas. However, the epsilon calculus is not widely used in the literature, and an important reason for this is its complex notation. One way to overcome this problem is to translate the epsilon calculus in a sequent calculus format. In this talk we will present possible translations of the epsilon formalism into a sequent calculus format and discuss their advantages and problems.

3.10 A short note on expressiveness and complexity in Propositional Dynamic Logic

Bruno Lopes (Fluminense Federal University – Rio de Janeiro, BR)

Joint work of Mario Benevides, Leandro Gomes, Bruno Lopes, Edward Hermann Heusler
Main reference Mario Benevides, Leandro Gomes, Bruno Lopes: "Towards determinism in PDL: relations and proof theory", Journal of Logic and Computation, p. exae022, 2024.
URL https://doi.org/10.1093/logcom/exae022

Propositional Dynamic logic is a sophisticated modal logic tailored to reason about programs. Several worth-of-studying fragments (controlling the complexity) and extensions (aiming to increase expressiveness) exist. We discuss some impacts of restricting to deterministic programs and how to enhance expressiveness using memory.

The first approach [1] relies on the usage of Guarded Kleene Algebras with Tests (GKAT – [2]) to define Strict Deterministic PDL programs (which will compose GPDL). As GKAT programs can be translated to Thompson's Automata and have its equivalence determined in almost linear time, the proposed algorithm is used to show that in GPDL $\models \langle\pi\rangle p \leftrightarrow \langle\eta\rangle p \Leftrightarrow \pi = \eta$.

A second approach consists on increasing the expressiveness by adding a memory to PDL [3]. Diferently from typical Memory Logics, the memory is in the syntax and not only in the model. We show that it is possible to reason about some context-sensitive languages and, for bounded memory, present a [factorial] translation to standard PDL.

References

1 Mario Benevides, Leandro Gomes, Bruno Lopes. *Towards determinism in PDL: relations and proof theory*. Journal of Logic and Computation, Oxford, UK, 2024
2 Steffen Smolka, Nate Foster, Justin Hsu, Tobias Kappé, Dexter Kozen, Alexandra Silva. *Guarded Kleene Algebras with Tests: verification of uninterpreted programs in nearly linear time*. Proceedings of the ACM on Programming Languages, Volume 4, Issue POPL, New York, USA, 2019

3 Mario Benevides, Bruno Lopes. *Memory Propositional Dynamic Logic*. In: Edward Hermann Haeusler, Luiz Carlos Pinheiro Dias Pereira, Jorge Petrucio Viana. (Org.). A Question is More Illuminating than an Answer. A Festschrift for Paulo A. S. Veloso. 1ed. College Publications, Oxford, UK, 2021

3.11 About Trust and Proof: An experimental framework for heterogeneous verification

Dale Miller (INRIA Saclay – Île-de-France, FR)

Joint work of Farah Al Wardani, Kaustuv Chaudhuri, Dale Miller
Main reference Farah Al Wardani, Kaustuv Chaudhuri, Dale Miller: "Formal Reasoning Using Distributed Assertions", in Proc. of the Frontiers of Combining Systems - 14th International Symposium, FroCoS 2023, Prague, Czech Republic, September 20-22, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14279, pp. 176–194, Springer, 2023.
URL https://doi.org/10.1007/978-3-031-43369-6_10

Information and opinions come to us daily from a wide range of actors, including scientists, journalists, and pundits. Some actors may be biased or malicious, while others rely on physical measurements, statistics, or in-depth research. Some sources may be signed or edited, while others are anonymous and unmoderated. Trusting information from such diverse sources is a serious challenge facing society today. In this paper, we will describe another domain – the world of machine-checked logic and mathematics – in which many similar issues can appear but in which tractable solutions are possible. Many actors (people or software systems) assert that certain logical statements are theorems in this domain. We describe the Distributed Assertion Management Framework (DAMF) that explicitly manages claims by theorem provers that they have proved certain theorems from associated contexts. Provers willing to trust other provers will be able to avoid rechecking proofs.

3.12 Second-order, well-behaved logic

Sara Negri (University of Genova, IT)

Joint work of Sara Negri, Matteo Tesi

The expressive capabilities of first-order logic are significantly expanded by second-order logic, which allows quantification over sets and properties. This extension addresses the natural demand in mathematics of expressing properties that involve quantification over all subsets or families of subsets of a given structure. Despite these advantages, full second-order logic lacks essential metalogical properties due to its impredicativity, which presents challenges in the development of proof systems. Our study tackles these challenges by introducing a G3-style calculus that admits the predicative comprehension schema, enabling a constructive cut elimination proof.

The calculi are extended to explore the proof theory of mathematical theories, with methods from first-order calculi being adapted, and structural results being established for both classical and intuitionistic calculi. Additionally, extensional equality and apartness are defined in second-order logic, showcasing the reduction of mathematical notions to pure logic. As an example, the theory of predicative second-order arithmetic is presented, and a variant of Herbrand's theorem tailored for predicative second-order intuitionistic logic is established, leading to the conservativity of predicative second-order Heyting arithmetic over its first-order counterpart. Furthermore, the interpolation theorem and the modal embedding of intuitionistic logic are extended to predicative second-order logic.

3.13 A rewriting logic approach to specification, proof-search, and meta-proofs in sequent systems

Carlos Olarte (Université Sorbonne Paris Nord – Villetaneuse, FR), Elaine Pimentel (University College London, GB)

Joint work of Carlos Olarte, Elaine Pimentel, Camilo Rocha
Main reference Carlos Olarte, Elaine Pimentel, Camilo Rocha: "A rewriting logic approach to specification, proof-search, and meta-proofs in sequent systems", J. Log. Algebraic Methods Program., Vol. 130, p. 100827, 2023.
URL https://doi.org/10.1016/J.JLAMP.2022.100827

In this talk, we demonstrate how inductive properties of propositional sequent systems can be automatically proved using the rewriting logic framework and its implementation in Maude. The properties of interest include the admissibility of weakening and contraction, rule invertibility, cut-admissibility, and identity expansion. We present examples of the L-Framework tool (`https://carlosolarte.github.io/L-framework/`) applied to various logical systems, including single-conclusion and multi-conclusion intuitionistic logic, classical logic, classical linear logic (and its dyadic system), intuitionistic linear logic, and normal modal logics.

3.14 Is, Ought, and cut

Edi Pavlovic (LMU München, DE)

Joint work of Norbert Gratzl, Edi Pavlovic
Main reference Norbert Gratzl, Edi Pavlovic: "Is, Ought, and Cut", J. Philos. Log., Vol. 52(4), pp. 1149–1169, 2023.
URL https://doi.org/10.1007/S10992-023-09701-Y

We demonstrate how to use proof-theoretic methods, specifically for a series of philosophically motivated deontic logics, to show that the principle of not deriving an "ought" from an "is" holds.

3.15 Proof Methods for Logics with Numerical Quantification

Ian Pratt-Hartmann (University of Manchester, GB)

In this talk, I ask about the limitations on various types of deductive systems for logics featuring numerical quantification. By "numerical quantification" I understand the use of constructions such as "There exist at most/at least/exactly n x such that ...", where

n is a natural number. It is known that many familiar fragments of first-order logic for which satisfiability is decidable can be extended with numerical quantification without losing decidability of satisfiability. At issue in this talk is what the decision methods in question have to look like. I begin with a negative result concerning the extension of the Aristotelian syllogistic with numerical quantification: I show that there exist no sound and complete collections of Aristotelian-like syllogisms for such logics. I then switch to the extension of the two-variable fragment of first-order logic with numerical quantification: I present a high-level account of decision procedures for this logic based on reduction to integer linear programming. Finally, I turn my attention to the one-variable fragment of first-order logic extended with numerical quantification: I ask whether there is a numerical extension of propositional resolution which is sound and complete for this logic.

3.16 Ultimate Glivenko?

Peter M. Schuster (University of Verona, IT)

Joint work of Sara Negri, Giulio Fellin, Peter Schuster

A simple criterion for the validity of an abstract Glivenko master theorem facilitates to instantiate the latter not only in conventional contexts such as Kuroda's but always when an axiom, e.g. stability, is added to a basic provability relation.

3.17 What is proof theory?

Lutz Straßburger (INRIA Saclay – Île-de-France, FR)

In this short talk I will discuss the problem of proof identity and explain how it is related to Hilbert's 24^{th} problem. I will also argue that not knowing when two proofs are "the same" has embarrassing consequences not only for proof theory but also for certain areas of computer science where formal proofs play a fundamental role, in particular the formal verification of software.

3.18 Negation inconsistent logics

Heinrich Wansing (Ruhr-Universität Bochum, DE)

Main reference Heinrich Wansing: "Beyond paraconsistency. A plea for a radical breach with the Aristotelean orthodoxy in logic", in: *Walter Carnielli on Reasoning, Paraconsistency, and Probability*, A. Rodrigues, H. Antunes, and A. Freire (eds), 2022, Springer, to appear 2025.

The talk contains material from a programmatic paper written in the context of an ERC grant, ERC-2020-ADG, ConLog, Contradictory Logics: A Radical Challenge to Logical Orthodoxy, H. Wansing, Beyond paraconsistency. A plea for a radical breach with the Aristotelean orthodoxy in logic, in: *Walter Carnielli on Reasoning, Paraconsistency, and Probability*, A. Rodrigues, H. Antunes, and A. Freire (eds), 2022, Springer, to appear 2024, and some comments on Gentzen-style proof systems for a certain non-trivial negation inconsistent logic.

After some terminological preliminaries on the notions of contradiction, inconsistency, and negation, I will make comments on paraconsistency, classical logic, and consistency. Then I will go beyond paraconsistency and will present a list of non-trivial negation inconsistent logics. Finally, I will introduce the non-trivial negation inconsistent connexive logic **C** and bilateral proof systems for **C**.

The references below may be considered *selected references* on this topic.

References

1. Ofer Arieli and Arnon Avron, Reasoning with logical bilattices, *Journal of Logic, Language and Information* 5 (1996), 25–63.
2. Newton da Costa, On the theory of inconsistent formal systems, *Notre Dame Journal of Formal Logic* 15 (1974), 497–510.
3. Allen Hazen and Jeffry Pelletier, Second-order logic of paradox, *Notre Dame Journal Formal Logic* (2018), 547–558.
4. Stanisław Jaśkowski, Rachunek zdań dla systemów dedukcyjnych sprzecznych, *Studia Societatis Scientiarum Torunensis*, Sectio A, Vol. I, No. 5, Toruń 1948, 57–77. English translation as “Propositional calculus for contradictory deductive systems”, *Studia Logica* 24 (1969), 143–157, and as “A propositional calculus for inconsistent deductive systems”, *Logic and Logical Philosophy* 7 (1999), 35–56. translation.
5. Norihiro Kamide, Paraconsistent double negation as classical and intuitionistic negations, *Studia Logica* 105 (2017), 1167–1191.
6. Norihiro Kamide and Heinrich Wansing, Proof theory of Nelson’s paraconsistent logic: A uniform perspective, *Theoretical Computer Science* 415 (2012), 1–38.
7. Satoru Niki, Double negation as minimal negation, *Journal of Logic, Language and Information* 32 (2023), 861–886.
8. Hitoshi Omori and Heinrich Wansing, On contra-classical variants of Nelson logic N4 and its classical extension, *Review of Symbolic Logic* 11 (2018), 805-820.
9. Hitoshi Omori and Heinrich Wansing, Contra-classicality in view of Dunn semantics, in: Katalin Bimbó (ed.), *Relevance Logics and other Tools for Reasoning. Essays in Honour of Michael Dunn*, College Publications, London, 2022,
10. Francesco Paoli, Logic and groups, *Logic and Logical Philosophy* 9 (2001), 109–128.
11. Graham Priest, Paraconsistent logic, in: D. Gabbay and F. Guenthner (eds), *Handbook of Philosophical Logic*, 2nd edition, vol. 6, Kluwer, Dordrecht, 2002, 287-393.
12. Heinrich Wansing, Falsification, natural deduction, and bi-intuitionistic logic, *Journal of Logic and Computation* 26 (2016), 425–450, published online July 2013.
13. Heinrich Wansing, Connexive logic, in: The Stanford Encyclopedia of Philosophy (Summer 2023 Edition) Edward N. Zalta (ed.), `http://plato.stanford.edu/archives/sum2023/entries/logic-connexive/`.
14. Heinrich Wansing, One heresy and one orthodoxy. On dialetheism, dimathematism, and the non-normativity of logic, *Erkenntnis* 89 (2024), 181–205.
15. Yale Weiss, Semantics for pure theories of connexive implication, *The Review of Symbolic Logic*, 15 (2022), 611–627.

3.19 The Epsilon Calculus in Non-classical Logics: Recent Results and Open Questions

Richard Zach (University of Calgary, CA)

Joint work of Matthias Baaz, Richard Zach
Main reference Matthias Baaz, Richard Zach: "Epsilon theorems in Intermediate Logics", J. Symb. Log., Vol. 87(2), pp. 682–720, 2022.
URL https://doi.org/10.1017/JSL.2021.103

The epsilon operator [1, 3] is mainly studied in the context of classical logic. It is a term forming operator: if $A(x)$ is a formula, then $\varepsilon x\, A(x)$ is a term – intuitively, a witness for $A(x)$ if one exists, but arbitrary otherwise. Its dual $\tau x\, A(x)$ is a counterexample to $A(x)$ if one exists. Classically, it can be defined as $\varepsilon x\, \neg A(x)$. Epsilon and tau terms allow the classical quantifiers to be defined: $\exists x\, A(x)$ as $A(\varepsilon x\, A(x))$ and $\forall x\, A(x)$ as $A(\tau x\, A(x))$.

Epsilon operators are closely related to Skolem functions, and the fundamental so-called epsilon theorems to Herbrand's theorem. Recent work with Matthias Baaz [2] investigates the proof theory of $\varepsilon\tau$-calculi in superintuitionistic logics. In contrast to the classical ε-calculus, the addition of ε- and τ-operators to intuitionistic and intermediate logics is not conservative, and the epsilon theorems hold only in special cases. However, it is conservative as far as the propositional fragment is concerned.

Despite these results, the proof theory and semantics of $\varepsilon\tau$-systems on the basis of non-classical logics remains underexplored.

References

1 Avigad, J., Zach, R.: The epsilon calculus. In: Zalta, E.N. (ed.) Stanford Encyclopedia of Philosophy. Fall 2024 edn. (2024), `https://plato.stanford.edu/entries/epsilon-calculus/`
2 Baaz, M., Zach, R.: Epsilon theorems in intermediate logics. The Journal of Symbolic Logic **87**(2), 682–720 (2022). https://doi.org/10.1017/jsl.2021.103
3 Zach, R.: Semantics and proof theory of the epsilon calculus. In: Ghosh, S., Prasad, S. (eds.) Logic and Its Applications. ICLA 2017, pp. 27–47. No. 10119 in LNCS, Springer, Berlin (2017). https://doi.org/10.1007/978-3-662-54069-5_4

4 Panel discussions

4.1 Special Session: Proof-theoretic Semantics

Alexander Gheorghiu (University College London, GB), Sara Ayhan (Ruhr-Universität Bochum, DE), and Victor Nascimento (Universidade do Estado do Rio de Janeiro, BR)

In *model-theoretic semantics* (M-tS), logical consequence is defined in terms of models; that is, abstract mathematical structures in which propositions are interpreted and their truth is judged. This includes, in particular, denotational semantics and Tarski's conception of logical consequence: a proposition ϕ follows model-theoretically from a context Γ iff every model of Γ is a model of ϕ,

$$\Gamma \models \phi \quad \text{iff} \quad \text{for all models } \mathcal{M}, \text{ if } \mathcal{M} \models \psi \text{ for all } \psi \in \Gamma, \text{ then } \mathcal{M} \models \phi$$

Proof-theoretic semantics (P-tS) [16] is an alternative approach to meaning and validity in which they are characterized in terms of *proofs* – understood as objects denoting collections of acceptable inferences from accepted premisses. It also concerns the semantics *of* proofs, understood as "valid" arguments.

To be clear, P-tS is not about providing a proof system. As Schroeder-Heister [15] observes, since no formal system is fixed (only notions of inference) the relationship between semantics and provability remains the same as it has always been: soundness and completeness are desirable features of formal systems.

The semantic paradigm supporting P-tS is *inferentialism* – the view that meaning (or validity) arises from rules of inference (see Brandom [2]). This may be viewed as a particular instantiation of the *meaning-as-use* paradigm by Wittgenstein [20] in which "use" in logic is understood as inferential rôle.

Heuristically, what differs is that (pre-logical) *proofs* in P-tS serve the rôle of *truth* in M-tS. This shift has substantial and subtle mathematical and conceptional consequences, as discussed below.

To illustrate the paradigmatic shift from M-tS to P-tS, consider the proposition "Tammy is a vixen". What does it mean? Intuitively, it means, somehow, "'Tammy is female' and 'Tammy is a fox'". On inferentialism, its meaning is given by the rules,

$$\frac{\text{Tammy is a fox} \quad \text{Tammy is female}}{\text{Tammy is a vixen}} \qquad \frac{\text{Tammy is a vixen}}{\text{Tammy is female}} \qquad \frac{\text{Tammy is a vixen}}{\text{Tammy is a fox}}$$

These merit comparison with the laws governing conjunction ($\wedge$), which justify the sense in which the above proposition is a conjunction,

$$\frac{\phi \qquad \psi}{\phi \wedge \psi} \qquad \frac{\phi \wedge \psi}{\phi} \qquad \frac{\phi \wedge \psi}{\psi}$$

There are several branches of research within P-tS – see, for example, the discussion on proof-theoretic validity in the Dummett-Prawitz tradition by Schroeder-Heister [14] – see also Gheorghiu and Pym [6]. Here, we concentrate on two topics: *base-extension semantics* and *bilateralism*.

Before proceeding to the topic details, we outline some important questions for the development of P-tS:

- *To what extent should P-tS depend on paradigms of proof?* On the one hand, different logics are more naturally expressed in some format of proofs than others (e.g., substructural logics typically favour sequent presentations more than natural deduction) and their P-tS may be influenced by this bias. Moreover, P-tS gives up an opportunity to challenge the current foundations and received dogma of the very concept of "proof" in logic. On the other hand, semantics ought to be syntax independent (in some sense). This may mean that a given notion of P-tS should be instantiable to different paradigms of proof, if none is taken as conceptually prior to the others (e.g., one may view a sequent calculus as providing "constructions" and natural deduction as providing the genuine article).
- *What might we expect of the relationship to M-tS?* Since M-tS is a powerful way of looking at logics, one may strive to show that the usual properties of M-tS are not lost if one transitions to P-tS. In particular, one may desire that the behaviour of models be represented in P-tS in some way that remains to be made precise. On the contrary, P-tS may offer an entirely different meta-theory that gives access to entirely distinct understandings of logics while forbidding other, perhaps useful, features of their extant semantics.

- *What is the real value of P-tS?* Developing the last point in a particular direction, we should consider what mathematical and computational value P-tS holds beyond its philosophical significance. To this end, one may begin by investigating how meta-theoretic properties of logics (e.g., compactness, categoricity, decidability, and so on) may be proved from the point of view of P-tS.

There are, of course, many more questions that one could ask. For example, P-tS may lead us to consider entirely new logics that have an obscure M-tS (if they have one at all). We defer further discussion of these matters to another time.

Base-extension Semantics

This section is concerned with a formalism in P-tS called *base-extension semantics* (B-eS). It follows the tradition of Piecha et al. [9] and Sandqvist [13].

The idea of B-eS begins with the notion of an *atomic system.* An atomic system is a collection of inferential relationships between *atoms.* They represent some beliefs that an agent may posses about the inferential relationship between thoughts. Piecha and Schroeder-Heister [17] and Sandqvist have given an analysis of them based on earlier work by Prawitz [10] and Schroeder-Heister.

Presently, we shall consider three types of atomic rules. Let $C, P_1, \ldots, P_n$ be atoms and $\mathbb{P}_1, \ldots, \mathbb{P}_n$ be finite, possibly empty, sets of atoms. The following are zero-, first-, and second-level atomic rules, respectively

$$\frac{}{C} \qquad \frac{P_1 \quad \ldots \quad P_n}{C} \qquad \frac{\begin{matrix}[\mathbb{P}_1] \\ P_1\end{matrix} \quad \ldots \quad \begin{matrix}[\mathbb{P}_n] \\ P_n\end{matrix}}{C}$$

The rules governing Tammy and her vixenhood above are atomic rules; specifically, they are *first-level* rules. Sandqvist [13] provides the following example of a second-level rule:

$$\frac{\text{A is a sibling of B} \quad \begin{matrix}[\text{A is a brother of B}] \\ P\end{matrix} \quad \begin{matrix}[\text{A is a sister of B}] \\ P\end{matrix}}{P}$$

Whether atomic rules correspond to "knowledge" or "definition" is a debated topic.

Atomic rules are read essentially as natural deduction rules in the sense of Gentzen. However, they are taken *per se* so that no substation is allowed. Thus, they are intuitively related to hereditary Harrop formulae in the sense of Miller.

A collection of atomic rules is an *atomic system.* We may restrict attention to certain atomic systems, in which case we call them *bases* ($\mathcal{B}$). Their reading as natural deduction rules (without substitution) determines a notion of *derivability in a base* ($\vdash_{\mathcal{B}}$).

Relative to a notion of derivability in a base ($\vdash_{\mathcal{B}}$), a B-eS is determined by a judgement called *support* ($\Vdash_{\mathcal{B}}$) defined inductively according to the structure of formulae with the base case (i.e., the support of atoms) given by *provability in a base.* This induces a validity judgement by quantifying our bases,

$$\Gamma \Vdash \phi \qquad \text{iff} \qquad \Gamma \Vdash_{\mathcal{B}} \phi \text{ for any base } \mathcal{B}$$

We illustrate this idea below.

Define a base $\mathcal{B}$ to be an atomic system that only contains zero- and first-level rules,

$$\frac{}{C} \qquad \frac{P_1 \quad \dots \quad P_n}{C}$$

Let $\overline{\mathbb{A}}$ denote the set of closed atoms and $\overline{\mathbb{T}}$ denote the set of closed terms. Relative to this notion of base, define a support relation ($\Vdash$) as follows:

$$\begin{array}{llll} \Vdash_{\mathcal{B}} P & \text{iff} & \vdash_{\mathcal{B}} P & \text{(At)} \\ \Vdash_{\mathcal{B}} \bot & \text{iff} & \Vdash_{\mathcal{B}} P \text{ for any } P \in \overline{\mathbb{A}} & (\bot) \\ \Vdash_{\mathcal{B}} \phi \wedge \psi & \text{iff} & \Vdash_{\mathcal{B}} \phi \text{ and } \Vdash_{\mathcal{B}} \psi & (\wedge) \\ \Vdash_{\mathcal{B}} \phi \to \psi & \text{iff} & \phi \Vdash_{\mathcal{B}} \psi & (\to) \\ \Vdash_{\mathcal{B}} \forall x \phi & \text{iff} & \Vdash_{\mathcal{B}} \phi[x \mapsto t] \text{ for any } t \in \overline{\mathbb{T}} & (\forall) \\ \Gamma \Vdash_{\mathcal{B}} \phi & \text{iff} & \text{for any } \mathcal{C} \supseteq \mathcal{B}, \text{ if } \Vdash_{\mathcal{C}} \psi \text{ for } \psi \in \Gamma, \text{ then } \Vdash_{\mathcal{C}} \phi & \text{(Inf)} \end{array}$$

Sandqvist (see also Makinson [7]) have shown that this characterises classical logic; that is,

$$\Gamma \Vdash \phi \quad \text{iff} \quad \phi \text{ follows classically from } \Gamma$$

Interestingly, $\Gamma \Vdash \phi$ is equivalent to $\Gamma \Vdash_{\emptyset} \phi$, suggesting that logical validity corresponds to *analytic* knowledge.

To express intuitionistic logic, we require extending the language with disjunction ($\vee$) and the existential quantifier ($\exists$). To this end, we may propose the following clauses:

$$\begin{array}{llll} \Vdash_{\mathcal{B}} \phi \vee \psi & \text{iff} & \Vdash_{\mathcal{B}} \phi \text{ or } \Vdash_{\mathcal{B}} \phi & (\vee) \\ \Vdash_{\mathcal{B}} \bot & \text{iff} & \Vdash_{\mathcal{B}} P \text{ for any } P \in \overline{\mathbb{A}} & (\exists) \end{array}$$

Piecha et al. [9] have shown that, surprisingly, intuitionistic logic is *incomplete* for this semantics. Subsequently, Stafford [18] showed that, in the propositional case, it corresponds to an intermediate logic known as *(general) inquisitive logic*.

We now observe that in the B-eS above, absurdity ($\bot$) is defined by *ex falso quodlibet*. This is quite unlike its treatment in more traditional M-tS. A philosophical motivation for this clause has been given by Dummett [3]. Following this motivation, Sandqvist [13] suggests the following alternative clauses:

$$\begin{array}{llll} \Vdash_{\mathcal{B}} \phi \vee \psi & \text{iff} & \text{for any } \mathcal{C} \supseteq \mathcal{B} \text{ and } P \in \overline{\mathbb{A}}, & \\ & & \text{if } \phi \Vdash_{\mathcal{C}} P \text{ and } \psi \Vdash_{\mathcal{C}} P, \text{ then } \Vdash_{\mathcal{C}} P & (\vee) \\ \Vdash_{\mathcal{B}} \exists x \phi & \text{iff} & \text{for any } \mathcal{C} \supseteq \mathcal{B} \text{ and } P \in \overline{\mathbb{A}}, & \\ & & \text{if } \phi[x \mapsto t] \Vdash_{\mathcal{C}} P \text{ for any } t \in \overline{\mathbb{T}}, \text{ then } \Vdash_{\mathcal{C}} P & (\exists) \end{array}$$

Here $\phi[x \mapsto t]$ is the result of replacing every free occurrence of x in ϕ by t.

To capture intuitionistic logic, some modification must be required at this point. To see this, consider Peirce's Law, $((P \to Q) \to P) \to P$. This formula is classically but not intuitionistically valid. Since it only contains implications and atoms, it is valid in the B-eS before the clauses for disjunction ($\vee$) and existential quantifier ($\exists$) were added, but that corresponds to classical logic. Hence, this intuitionsitic logic is not complete for this B-eS.

We require only a small but significant change for intuitionistic logic: we now permit second-level rules in bases,

$$\frac{\begin{matrix}[\mathbb{A}_1] \\ P_1\end{matrix} \quad \dots \quad \begin{matrix}[\mathbb{A}_n] \\ P_n\end{matrix}}{c}$$

Sandqvist (see also Gheorghiu) have shown that the result indeed corresponds to intuitionistic logic,

$$\Gamma \Vdash \phi \qquad \text{iff} \qquad \phi \text{ follows intuitionistically from } \Gamma$$

Though B-eS appears to be closely related to M-tS (esp. possible world semantics in the sense of Beth and Kripke), the formal connection remains an enigma. Indeed, while Makinson [7] (resp. Eckhardt and Pym [4]) have made formal connections between the M-tS and B-eS of classical logic (resp. normal modal logic), the analogous connections for intuitionistic logics are currently unknown. Part of the challenge is in the considerably different ways that disjunctive structures (i.e., $\bot$, $\vee$, $\exists$) are treated.

The above work on the B-eS of classical and intuitionistic logic has been extended by Eckhardt and Pym [4] to modal logic, by Gheorghiu et al. [5] and Buzoku to substructural logic (namely, *intuitionistic Linear Logic* and *the logic of Bunched Implications*), and by Nascimento et al. to ecumenical logic. Closely related approaches have also been developed by Goldfarb and Nascimento and Stafford [8].

Bilteralism

Logical bilateralism can be very generally described as an approach to meaning and consequence on the grounds of a symmetry between certain notions, like assertion and denial, proof and refutation or truth and falsity, in that both are taken as primitive and not, as in conventional "unilateralist" approaches, merely reducing the latter to the former, more primary notion. In recent years, the field of logical bilateralism has seen significant development with various systems being developed that showcase a range of orientations within this framework. In Rumfitt's seminal paper [12], in which the term "bilateralism" was introduced, he means to give a motivation for how the natural deduction rules of classical logic lay down the meaning of the connectives once we consider a calculus containing introduction and elimination rules determining not only the assertion conditions for the connectives but also the denial conditions.

This is realized by using signed formulas in the form of "+A" and "–A" where "+" and "–" are used as force indicators. Smiley developed a similar approach and there are also other and earlier works promoting general bilateralist ideas. While several works explore and refine this approach to bilateralism in that the main focus is on natural deduction style proof systems with assertion and denial conditions, there have been developments in other directions, in which bilateralist considerations play an equally central role. Some propose a new way of reading a (classical) sequent calculus with multiple conclusions, namely by way of defining an inference, represented by a sequent, as valid if and only if it is incoherent to assert all the premises (i.e., the formulas on the left side of the sequent sign), while simultaneously denying all the conclusions (i.e., the formulas on the right side of the sequent sign) – see, for example, Restall [11]. Here, the bilateralist considerations do not arise in the design of a distinctive proof system, but in the interpretation of an already existing proof calculus by way of taking assertion and denial as dual notions.

The approach presented in the special session focuses not so much on the speech acts of assertion and denial but on a duality between different inferential relationships, which in turn give rise to motivating proof systems with dual derivability relations. Such proof systems displaying provability and refutability can be represented both in natural deduction and in sequent calculus style (see, for example, Wansing [19] and Ayhan [1]). On such a view it can be asked, then, how these dual derivability relations can be implemented on a meta-level. In a sequent calculus setting, for example, this would mean not only to have signed sequents, displaying provability and refutability within sequents, but also displaying the dual relations between sequents.

References

1. S. Ayhan. Meaning and Identity of Proofs in a Bilateralist Setting: A two-sorted typed lambda-calculus for proofs and refutations. *Journal of Logic and Computation*, forthcoming.
2. R. Brandom. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, 2000.
3. M. Dummett. *The Logical Basis of Metaphysics*. Harvard University Press, 1991.
4. T. Eckhardt and D. J. Pym. Base-extension Semantics for Modal Logic. *Logic Journal of the IGPL*, 2024. doi:10.1093/jigpal/jzae004.
5. A. V. Gheorghiu, T. Gu, and D. J. Pym. Proof-theoretic Semantics for Intuitionistic Multiplicative Linear Logic. In R. Ramanayake and J. Urban, editors, *Automated Reasoning with Analytic Tableaux and Related Methods – TABLEAUX*, pages 367–385. Springer, 2023.
6. A. V. Gheorghiu and D. J. Pym. From Proof-theoretic Validity to Base-extension Semantics for Intuitionistic Propositional Logic *Studia Logica*, 2024 (forthcoming)
7. D. Makinson. On an Inferential Semantics for Classical Logic. *Logic Journal of IGPL*, 22(1):147–154, 2014.
8. V. Nascimento. *Foundational Studies in Proof-theoretic Semantics*. PhD thesis, Universidade do Estado do Rio de Janeiro, 2023.
9. T. Piecha. Completeness in Proof-theoretic Semantics. In *Advances in Proof-theoretic Semantics*, pages 231–251. Springer, 2016.
10. D. Prawitz. *Natural Deduction: A Proof-theoretical Study*. Dover Publications, 2006 [1965].
11. G. Restall. Multiple conclusions. In P. Hájek, L. Valdes-Villanueva, and D. Westerståhl, editors, *Logic, Methodology and Philosophy of Science: Proceedings of the Twelfth International Congress*, pages 189–205. King's College Publications, London, 2005.
12. I. Rumfitt. "Yes" and "No". *Mind*, 109(436):781–823, 2000.
13. T. Sandqvist. Base-extension Semantics for Intuitionistic Sentential Logic. *Logic Journal of the IGPL*, 23(5):719–731, 2015.
14. P. Schroeder-Heister. Validity Concepts in Proof-theoretic Semantics. *Synthese*, 148(3):525–571, 2006.
15. P. Schroeder-Heister. Proof-Theoretic versus Model-Theoretic Consequence. In M. Pelis, editor, *The Logica Yearbook 2007*. Filosofia, 2008.
16. P. Schroeder-Heister. Proof-Theoretic Semantics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2018 edition, 2018.
17. P. Schroeder-Heister and T. Piecha. Atomic Systems in Proof-Theoretic Semantics: Two Approaches. In Ángel Nepomuceno Fernández, O. P. Martins, and J. Redmond, editors, *Epistemology, Knowledge and the Impact of Interaction*. Springer Verlag, 2016.
18. W. Stafford. Proof-Theoretic Semantics and Inquisitive Logic. *Journal of Philosophical Logic*, 2021.
19. H. Wansing. A more general general proof theory. *Journal of Applied Logic*, 25:23–46, 2017.
20. L. Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 2009.

4.2 Special Session: Universal Proof Theory

Raheleh Jalali (The Czech Academy of Sciences – Prague, CZ), Timo Lang (University College London, GB), and Iris van der Giessen (University of Birmingham, GB)

Universal Proof Theory is a recent research program that aims to study the mathematical properties of proof systems in a generic manner (analogous to the generic study of algebras in Universal Algebra). The term Universal Proof Theory was first introduced in preprint [1]. The program consists of three fundamental problems: (1) the existence problem to investigate the existence of certain kinds of proof systems for a given logic, (2) the equivalence problem to study the natural notions of equivalence of proof systems, and (3) the characterization problem to investigate the possible characterizations of proof systems. In the session, we provide an overview of the results which are mainly achieved for the existence problem (i.e., [3, 1, 4, 5, 2]). We discuss a small unsolved problem to provide a flavor of what this research involves. We invite the audience to connect their work to this program so that we create a better understanding of Universal Proof Theory within the broader community of Proof Theory.

References

1 A. Akbar Tabatabai and R. Jalali. *Universal proof theory: Semi-analytic rules and Craig Interpolation*. Annals of Pure and Applied Logic, 2024, https://doi.org/10.1016/j.apal.2024.103509
2 A. Ciabattoni, N. Galatos and K. Terui. *Algebraic proof theory for substructural logics: cut-elimination and completions*. Annals of Pure and Applied Logic, 163(3):266–290, 2012
3 R. Iemhoff. *Uniform interpolation and the existence of sequent calculi*. Annals of Pure and Applied Logic, 170(11):102711 , 2019
4 F.M. Lauridsen. *Intermediate logics admitting a structural hypersequent calculus*. Studia Logica, 107:247–282, 2019
5 S. Negri and J. von Plato. *Cut elimination in the presence of axioms*. The Bulletin of Symbolic Logic, 4(4):418-435, 1998

4.3 Special Session: Circular and Non-wellfounded Proofs

Alexis Saurin (CNRS – Paris, FR), Anupam Das (University of Birmingham, GB), and Abhishek De

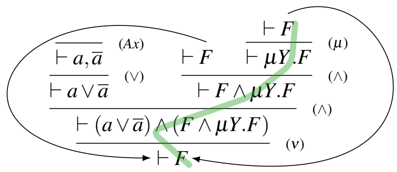
Special session held on 22nd August 2024
Contributors: Anupam DAS, Abhishek DE & Alexis SAURIN

Non-wellfounded proofs and their regular fragment – cyclic, or circular proofs – appear in a very natural way when studying inductive and coinductive reasoning [5, 8, 12, 13] (such as in variants of the μ-calculus, or in logics with inductively-defined predicates). While the complex inference rules for induction and coinduction are replaced by simpler fixed-point unfolding rules, the existence of infinite branches in proof trees requires that a global condition is imposed on derivations in order to ensure soundness and to allow a structured proof-theoretical analysis as shown below:

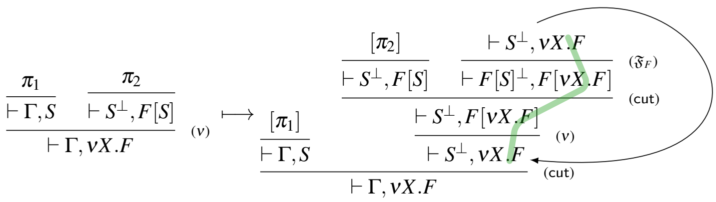
$$\dfrac{\dfrac{\dfrac{\vdots}{\vdash \mu X.X}\,(\mu)}{\vdash \mu X.X}\,(\mu) \qquad \dfrac{\dfrac{\vdots}{\vdash \nu X.X,\Gamma}\,(\nu)}{\vdash \nu X.X,\Gamma}\,(\nu)}{\vdash \Gamma}\,(\mathsf{cut})$$

The special session presented some complements on Bahareh AFSHARI's invited talk along the following lines and with a specific focus on the Curry-Howard correspondence between proofs and programs.

- Abhishek DE presented the logical frameworks under study as well as the thread progress-condition for non-wellfounded and regular proofs (illustrated below, with $F = \nu X.(a \vee \overline{a}) \wedge (X \wedge \mu Y.X)$) and discussed its decidability and complexity. He also explained how the condition ensures the soundness of the proof system [5, 7, 8, 12].



- Anupam DAS compared finitary (co)induction inference rules [1, 5] and the provability in cyclic proofs [5, 8]: he showed that, under a very general condition, circular proofs subsume finitary induction and coinduction [8] (see figure below) and reviewed the literature studying the converse property [3, 4, 8, 11].



- Finally, Alexis SAURIN reviewed cut-elimination theorems for those proof systems [2, 8, 9, 10] including non-wellfounded derivations and the challenge raised by the non-wellfounded proof trees. On the way, he showed, on some examples, how inductive and coinductive data can be represented as non-wellfounded and circular proofs [6, 8].

References

1 David Baelde: Least and Greatest Fixed Points in Linear Logic. ACM Trans. Comput. Log. 13(1): 2:1-2:44 (2012)

2 David Baelde, Amina Doumane, Alexis Saurin: Infinitary Proof Theory: the Multiplicative Additive Case. CSL 2016: 42:1-42:17

3 Stefano Berardi, Makoto Tatsuta: Classical System of Martin-Löf's Inductive Definitions Is Not Equivalent to Cyclic Proof System. FoSSaCS 2017: 301-317

4 Stefano Berardi, Makoto Tatsuta: Equivalence of inductive definitions and cyclic proofs under arithmetic. LICS 2017: 1-12

5 James Brotherston, Alex Simpson: Sequent calculi for induction and infinite descent. J. Log. Comput. 21(6): 1177-1216 (2011)
6 Gianluca Curzi, Anupam Das: Computational expressivity of (circular) proofs with fixed points. LICS 2023: 1-13
7 Anupam Das, Abhishek De, Alexis Saurin: Decision Problems for Linear Logic with Least and Greatest Fixed Points. FSCD 2022: 20:1-20:20
8 Amina Doumane: On the infinitary proof theory of logics with fixed points. (Théorie de la démonstration infinitaire pour les logiques à points fixes). Paris Diderot University, France, 2017
9 Jérôme Fortier, Luigi Santocanale: Cuts for circular proofs: semantics and cut-elimination. CSL 2013: 248-262
10 Alexis Saurin: A Linear Perspective on Cut-Elimination for Non-wellfounded Sequent Calculi with Least and Greatest Fixed-Points. TABLEAUX 2023: 203-222
11 Alex Simpson: Cyclic Arithmetic Is Equivalent to Peano Arithmetic. FoSSaCS 2017: 283-300
12 Christoph Sprenger, Mads Dam: On the Structure of Inductive Reasoning: Circular and Tree-Shaped Proofs in the µ-Calculus. FoSSaCS 2003: 425-440
13 Igor Walukiewicz: Completeness of Kozen's Axiomatisation of the Propositional µ-Calculus. Inf. Comput. 157(1-2): 142-182 (2000)

Participants

- Matteo Acclavio
University of Southern Denmark – Odense, DK
- Bahareh Afshari
University of Gothenburg, SE
- Sara Ayhan
Ruhr-Universität Bochum, DE
- Eben Blaisdell
University of Pennsylvania – Philadelphia, US
- Yll Buzoku
University College London, GB
- Kaustuv Chaudhuri
INRIA Saclay – Île-de-France, FR
- Zhibo Chen
Carnegie Mellon University – Pittsburgh, US
- Anupam Das
University of Birmingham, GB
- Abishek De
University of Birmingham, GB
- Amy Felty
University of Ottawa, CA
- Alexander Gheorghiu
University College London, GB
- Marianna Girlando
University of Amsterdam, NL
- Rajeev P. Gore
Turner, AU
- Tao Gu
University College London, GB
- Andrzej Indrzejczak
University of Lodz, PL
- Raheleh Jalali
The Czech Academy of Sciences – Prague, CZ
- Timo Lang
University College London, GB
- Anela Lolic
TU Wien, AT
- Bruno Lopes
Fluminense Federal University – Rio de Janeiro, BR
- Robin Martinot
Utrecht University, NL
- Dale Miller
INRIA Saclay – Île-de-France, FR
- Victor Nascimento
Universidade do Estado do Rio de Janeiro, BR
- Sara Negri
University of Genova, IT
- Carlos Olarte
Université Sorbonne Paris Nord – Villetaneuse, FR
- Edi Pavlovic
LMU München, DE
- Elaine Pimentel
University College London, GB
- Ian Pratt-Hartmann
University of Manchester, GB
- Revantha Ramanayake
University of Groningen, NL
- Alexis Saurin
CNRS – Paris, FR
- Peter M. Schuster
University of Verona, IT
- Sana Stojanovic-Djurdjevic
University of Belgrade, RS
- Lutz Straßburger
INRIA Saclay – Île-de-France, FR
- Iris van der Giessen
University of Birmingham, GB
- Fernando Velázquez Quesada
University of Bergen, NO
- Heinrich Wansing
Ruhr-Universität Bochum, DE
- Richard Zach
University of Calgary, CA

Report from Dagstuhl Seminar 24342

Leveraging AI for Management Decision-Making

Stefan Feuerriegel*[1], Foster Provost*[2], and Galit Shmueli*[3]

1 **Ludwig-Maximilians-Universität München, DE.** feuerriegel@lmu.de
2 **New York University, US.** fprovost@stern.nyu.edu
3 **National Tsing Hua University – Hsinchu, TW.** galit.shmueli@iss.nthu.edu.tw

Abstract

Artificial intelligence (AI) is transforming decision-making across industries and management functions, which leads to increased operational efficiency and creates significant economic impact. A recent surge in attention to AI in business decision-making has been driven by new AI technologies – such as deep learning, causal machine learning, generative AI and explainable AI – and their applications in areas like operations, marketing, information systems, and quality management. Yet, the potential of AI to optimize business decisions also introduces ethical, legal, and societal challenges, particularly in high-stakes business settings. This motivates our Dagstuhl Seminar, which aimed to foster interdisciplinary collaboration between scholars in management and computer science, as well as practitioners from industry. As a result, the seminar generated new suggestions for the field to evolve in the future by identifying new research opportunities with managerial relevance.

Seminar August 18–21, 2024 – https://www.dagstuhl.de/24342
2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Social and professional topics → Computing and business; Computing methodologies → Machine learning; Information systems
Keywords and phrases applications, business, decision-making, management, marketing
Digital Object Identifier 10.4230/DagRep.14.8.24

1 Executive Summary

Stefan Feuerriegel
Foster Provost
Galit Shmueli

Introduction

Artificial intelligence (AI) has been named a core element of the "fourth industrial revolution" [40]. According to recent estimates by McKinsey & Company, AI has the potential to deliver the added global economic value of $13–20 trillion annually [5].

AI is increasingly being embraced for decision-making in management, both across a wide array of industries (e.g., healthcare, banking, education, manufacturing, retail) and functions (e.g., marketing, accounting, operations, IT). For example, AI can be used for modeling customer behavior [2, 6, 25, 26, 4]. These predictions can also serve as input for better decision-making. Examples include assortment optimization [24, 19], investment

* Editor / Organizer

decisions [32], scheduling [42], allocation decisions [29, 31, 46], and pricing [1]. AI can predict business failures and thus act as an early warning system for improving service quality [34]. AI can help to locate drivers of low quality and eventually improve product quality [44, 45].

Recent advancements in AI research hold great promise for decision-making in businesses and organizations. Driven by the surge in data availability, computing power, and algorithmic advancement, contemporary AI algorithms are capable of emulating human decision-making and judgment [39]. This places AI in a position to augment and automate a wide range of management decisions within companies and organizations. At the same time, the use of AI for decision making, especially in high-risk applications, poses ethical and legal challenges such as the use of algorithmic risk assessment tools in criminal justice [18]. Likewise, new implications arise from emerging AI acts (e.g., the EU AI Act) with crucial implications for how AI applications must be designed.

A key enabler for data-driven decision-making in business and organizations are new AI technologies [17]. For example, deep learning can empower better decisions in business analytics and operational decision-making [30, 28]. Causal machine learning (ML) allows for optimal targeting (e.g., of customer coupons) by estimating and subsequently leveraging individualized treatment effects [37, 20, 11, 14, 15]. Probabilistic machine learning fuses methods from a statistical foundation with flexible building blocks from neural networks to yield models that are both flexible and explainable for practitioners in risk management [38, 35, 36]. Further, explainable AI (XAI) has emerged as a principled, user-centered tool not only for explaining black-box prediction models but also for explaining the decisions that are made or recommended by AI systems [33, 45, 16, 27, 3], which can be used to identify root causes of bad quality and thereby inform better decision-making in quality management [44]. Likewise, generative AI [12] and AI fairness [9, 7, 10] offer new research opportunities. Importantly, the aforementioned examples can only be solved effectively through new AI technologies that have been developed in recent years. At the same time, the use of advanced AI on digital platforms, especially the marriage of reinforcement learning with behavior modification techniques, has spurred controversy, creating adverse effects to humans and societies, such as addiction, social discord, and political polarization [22]. Existing and envisioned combinations of prediction and causal behavior modification have implications to platforms and their business customers [23]. These technologies have also created new types of barriers for academic researchers [41, 21].

Aims of the seminar

The aim of our Dagstuhl Seminar "Leveraging AI for Management Decision-Making" (24342) was to discuss the future of research on AI/ML in businesses and organizations, and how the field should evolve. We especially sought to focus on "rethinking" the field by discussing the current state of AI/ML in businesses and organizations, discussing thought-provoking questions (e.g., is explainable AI really needed in practice? Where can generative AI actually lead to productivity gains? Does the algorithmic approach to fairness hold much for the future of AI ethics?), and maybe identifying and elevating new important research questions.

Our intended outcome was to reach a joint position (as a group) on what are important and unimportant research directions and what those directions should be going forward. What are the challenges? What are the opportunities? What research questions deserve attention? What questions are getting more attention than they really deserve? Below, we summarize our thoughts where we discuss existing research gaps and make suggestions for the field going forward.

Prior to the seminar, a survey was shared with all participants to identify key topics of interest around which we then designed our discussions. Most participants were primarily interested in topics related to method design and development, as well as the practical applications of AI. Some also expressed interest in evaluating these methods and understanding their impact on organizations. There was a strong focus on exploring the broader implications of AI, particularly in areas like ethics and governance. When asked about specific topics they would like to discuss during the seminar, participants showed the most interest in explainable AI, followed by causal ML, and generative AI. We eventually decided to prioritize the first two – explainable AI and causal ML – thus anticipating that generative AI will naturally arise in all sessions due to its prominence and thus regardless of whether it is a dedicated topic. Many indicated they were also keen to explore how AI can be applied effectively, with discussions centered around overcoming practical challenges in real-world deployments. Other popular topics include the ethics and governance of AI, its economic impact, and the implications for the future of work. However, there was less interest in topics like AI literacy and hybrid work environments. Further, participants also suggested additional topics for discussion. Some highlighted the importance of understanding the behavioral impacts of AI, such as long-term reliance on AI systems and the potential for deskilling human workers.

Organization of the seminar

We designed our Dagstuhl Seminar as an "un-conference". By following the format of an un-conference, we eliminated the traditional sequence of research presentations from our agenda. Instead, we aimed to focus on interactivity, collaboration, and co-creation, by making space for discussions of different forms regarding how to shape the field in the future. We held discussions with the full group as well as in smaller break-out groups, where subgroups changed from day to day; we obtained information from individuals via surveys; the schedule also encouraged informal one-on-one or small group discussions while socializing. These various modes of interaction were critical, because our seminar attracted participants from a diverse crowd, from academia and industry, from method research to behavioral research, from marketing to operations. Such diverse researchers typically do not meet or interact, and hence we seized the opportunity to foster novel interactions.

We aimed to learn from each other and create more impact. For example, we actively asked each participant in the get-to-know session to provide a summary statement about their current research and where they would like to go. Throughout the seminar, there were many opportunities to potentially start new collaborations. To spur discussion, we organized short "inspiration exchanges", which were designed as kick-offs to our breakout sessions. Hence, the idea was primarily to discuss the current state of research and point to gaps and needs to elicit forward-thinking. Here, we selected the topics prior to the seminar based on a survey that was sent to the participants. As a result, we identified two important breakout sessions: (1) AI and causality, and (2) AI and responsibility. We summarize the discussions and findings from both breakout sessions below.

2 Table of Contents

3 Summary of Breakout Session on "AI and Causality"

This breakout session focused on exploring the interface of AI and causality by discussing open questions and emerging challenges.

3.1 What is causal ML?

The session started with an inspirational short presentation introducing the concept of causal ML [43, 14]. Causal ML is a branch of machine learning focused on identifying and understanding cause-and-effect relationships, rather than simply detecting patterns in data. Traditional machine learning models excel at predicting outcomes by recognizing correlations, but they often struggle to determine whether a specific factor directly causes a change in the target variable. Causal ML attempts to address this limitation by combining statistical techniques from causal inference with modern machine learning approaches. It allows researchers to answer questions like "What will happen if we change this feature?" or "What is the impact of a particular intervention?" This is particularly useful in fields like management where the aim is to to understand how management decisions will influence outcomes, so that the best decision for the business can be made.

In practice, Causal ML uses traditional statistical and econometric methods such as randomized controlled trials, propensity score matching, and instrumental variables to isolate causal effects. Additionally, more advanced ML-based techniques like causal forests and doubly robust estimators have been developed to handle complex, high-dimensional data where traditional methods might fall short. By leveraging these approaches, Causal ML models can go beyond predictions to inform actionable insights, such as making recommendations for business decisions in strategy, marketing, etc. Generally, causal ML follows the same frameworks as traditional causal inference, but the use of ML changes the underlying estimation strategy and thus offers several benefits in business practice, such as making personalized decisions for customers or users. Moreover, the goals of causal inference for decision making can be different from traditional causal-effect estimation [14], motivating the application of alternative machine learning approaches [15].

3.2 Key discussions

Defining Concepts and Aligning Terminology:

The discussions highlighted the need for clearer definitions and a shared vocabulary in the field. Participants acknowledged that "causality" is often interpreted differently across disciplines, which can hinder effective collaboration among different fields and between practitioners and researchers. Aligning terminology and notation was seen as a critical step for advancing research and practical applications.

One of the provocative questions debated was whether prediction (causal or non-causal) is sufficient for effective decision-making or if more granular causal inference is necessary. Participants discussed scenarios where accurate predictions might suffice (e.g., short-term business decisions), versus contexts where a deeper understanding of causal relationships is crucial (e.g., long-term strategic planning or policy-making). The consensus was that while predictive models are valuable, they may fall short in areas where understanding the "why" behind outcomes is essential. Yet, beyond that, the participants see a large need for more research and eventually evidence allowing for decision support.

As mentioned above, one of the biggest challenges in causal ML is deciding when to prioritize causal models over traditional predictive models. Predictive ML models, such as those used for classification or regression, are primarily designed to optimize accuracy based on historical patterns, without necessarily understanding the underlying relationships. However, when the goal is to make decisions that could change the environment or influence outcomes (like deciding on a new marketing strategy or a treatment plan in healthcare), causal modeling could be the "go to" approach. Yet, the challenge lies in identifying these scenarios where the extra complexity and effort of causal analysis are justified. Deep causal understanding requires carefully constructed assumptions and a deep understanding of the data-generating process, which can be resource-intensive. On the other hand, causal prediction may be sufficient (and quite effective) for decision making – and is often overlooked [13]. Without clear indicators of when causal methods are needed, organizations may either waste resources on unnecessary complexity or, conversely, miss opportunities to derive actionable insights where causality matters.

From a practical perspective, adopting causal ML techniques is often hindered by non-technical challenges such as the availability of skilled personnel, time, and budget constraints. Traditional ML has broad and accessible toolchains, tutorials, and community support, facilitating rapid training and deployment of models. In contrast, using ML to achieve causal understanding requires additional specialized training, including knowledge of statistical theory and causal inference frameworks, which are not as widely understood or accessible. Moreover, the data requirements for causal models can be more stringent, often requiring richer datasets or careful experimental designs. Together, this may create barriers for companies trying to leverage causal insights, especially those with limited resources.

A recurring question was whether algorithms for automated decision-making always require a causal framework. Opinions varied, with some arguing that in certain applications (e.g., dynamic pricing, recommendation systems), causal understanding is not always necessary, while others stressed that understanding causal relationships is crucial for ensuring fairness and avoiding unintended consequences.

Participants raised concerns about the limitations of causal inference from observational data, particularly when confounding factors and model misspecifications are present. The discussion emphasized the need for developing more robust methodologies to derive reliable causal insights, especially in real-world applications where randomized experiments are often impractical.

3.3 Conclusion

The breakout sessions successfully brought to light the complexities and nuances involved in integrating causality with AI, especially around formal definitions, guidelines when causal modeling is beneficial (and when not), and effective approaches. The discussions highlighted the importance of interdisciplinary collaboration to advance research, develop practical tools, and address the societal implications of AI-driven decision-making.

4 Summary of Breakout Session on "AI and Responsibility"

The second breakout session was centered around responsibility in AI adoption. It brought together participants to discuss the ethical, social, and governance implications of AI technologies. The breakout session was intentionally named broadly, so that it would include specific methodologies (e.g., explainable AI) as well as organizational implications. Eventually, the conversation was broad, covering explainable AI, generative AI, algorithmic fairness, and the societal impact of AI systems. A central focus was on the challenges and pitfalls of ensuring accountability in AI systems, which require a certain level of transparency in how decision-making algorithms operate.

4.1 What is explainable AI?

The starting point was an inspirational high-level short introduction and exchange discussing methods for explaining AI algorithms, models, and decisions, as well as the underlying objectives, rationales, and limitations [8]. Explainable AI (XAI) refers to a broad set of techniques and methods used to make the decisions, predictions, and/or inner workings of AI models more transparent and understandable to humans. Many AI systems, especially those that use complex algorithms like deep learning, are often seen as "black boxes" because their decision-making processes are not easily interpretable. This lack of transparency can be problematic, especially in sensitive areas like finance, credit lending, or law enforcement, where understanding how decisions are made is crucial for trust and accountability. Explainable AI aims to bridge this gap by providing insights into how or why models reach their conclusions, which can help users trust and effectively use AI systems.

The goal of XAI is to provide explanations that are not only faithful but also understandable to different stakeholders, such as data scientists, business leaders, or end-users. For example, in quality management, an XAI model might highlight which processes contributed most to a low quality level, helping manufacturing make more informed decisions about where to improve production processes. Techniques like feature importance scores, counterfactual explanations, interpretable models such as rule lists or shallow decision trees, and visualization tools are commonly used to make AI models and decisions more interpretable. By making AI decisions clearer, explainable AI could – in principle – help ensure that systems are not only accurate but also fair, reliable, and aligned with ethical standards, especially when those decisions impact people's lives. However, as we discuss below, there was large dispute among the participants whether this promise is true in practice.

Despite its benefits, XAI suffers from several limitations. One major limitation in practice is that it is often unclear for whom any particular explanation method was designed, and there is often a mismatch between the capabilities of these methods and the needs in practice. This leads to frequent cases where explanation methods fail to serve the needs of users or stakeholders who would like to use such explanations for decision-making. For example, it is often unclear how certain methods such as feature importance can help decision-makers generate insights that are actionable and that can be translated into better decisions. Another major challenge is that, for highly complex models like deep neural networks, providing simple and intuitive explanations can be difficult. Sometimes, explanations generated by explainable AI methods may oversimplify the decision process, leading to misunderstandings or even incorrect conclusions. Finally, there is little consensus on what a "good" explanation is.

4.2 Key discussions

Participants emphasized the importance of XAI for fostering transparency and trust, particularly in sectors like healthcare, finance, and criminal justice. However, the group also critically examined the current state of XAI research and its use in practice, thereby emphasizing salient limitations. A recurring concern was that many so-called "explanations" provided by AI systems are overly simplified, often failing to capture the complexities of the underlying algorithm. Another problem is that the explanations are rarely aligned with the needs of stakeholders in practice [33], so that actions derived from XAI may lead to negative outcomes. Together, the discussions highlighted that XAI can create a false sense of understanding or accountability, particularly when explanations are designed more for compliance than genuine transparency.

The participants also see a large potential for future research and thus for how the field could move forward. The discussion highlighted the need for more rigorous evaluations of the impact of AI explanations on decision-makers, particularly in contexts where human lives or rights are at stake. Currently, the evaluations are rarely aligned with how XAI systems are used in practice. For example, many evaluations are often one-sided and only focus on the role of programmers, while more holistic evaluation approaches that account for the different roles of stakeholders are still scarce. Another point for future research is to clarify what we mean under 'understandable' as this is a necessary condition to make assessments as to when one model is more interpretable than another.

The session also touched on the rise of regulatory approaches to AI (e.g., some regulatory frameworks offer a right for explanations in the context of data-driven decision-making). Here, participants emphasized the importance of regulatory and governance frameworks to provide oversight for the development and deployment of AI to prevent harm while enabling innovation.

4.3 Conclusion

The session underscored the complexity of ensuring responsible AI in practice. While explainable AI is a step towards greater transparency, participants highlighted that algorithmic explanations alone are not enough. In response to the session, the participants aim to create a commentary that offers critical reflections and guidance. We plan to discuss the following directions in our commentary: Which method should be applied by whom, in what context, and with what goal in mind? We will thus take a step back and will propose a comprehensive framework that shows the translation of real-world business problems into the model world while highlighting the critical dimensions that influence the success of XAI initiatives.

References

1 Svenja Bergmann and Stefan Feuerriegel. Machine learning for predicting used car resale prices using granular vehicle equipment information. *Expert Systems with Applications*, 263:125640, 2024.

2 Lennart Baardman, Igor Levin, Georgia Perakis, and Divya Singhvi. Leveraging comparables for new product sales forecasting. *International Journal of Operations & Production Management*, 27(12):2340–2343, 2018.

3 Kevin Bauer, Moritz von Zahn, and Oliver Hinz. Expl(ai)ned: The impact of explainable artificial intelligence on cognitive processes. *Information Systems Research*, 2023. forthcoming.

4 Ruomeng Cui, Santiago Gallino, Antonio Moreno, and Dennis J Zhang. The operational value of social media information. *Production and Operations Management*, 27(10):1749–1769, 2018.
5 Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, and Kate Smaje. *The economic potential of generative AI*. McKinsey & Company, 2023.
6 Real Carbonneau, Kevin Laframboise, and Rustam Vahidov. Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184(3):1140–1154, 2008.
7 Maria De-Arteaga, Stefan Feuerriegel, and Maytal Saar-Tsechansky. Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, 31(10):3749–3770, 2022.
8 Koen W De Bock, Kristof Coussement, Arno De Caigny, Roman Słowiński, Bart Baesens, Robert N Boute, Tsan-Ming Choi, Dursun Delen, Mathias Kraus, Stefan Lessmann, et al. Explainable ai for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, 317(2):249–272, 2024.
9 Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. A sociotechnical view of algorithmic fairness. *Information Systems Journal*, 32(4):754–818, 2022.
10 Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe. Fair ai: Challenges and opportunities. *Business & information systems engineering*, 62:379–384, 2020.
11 Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30:958–968, 2024.
12 Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative AI. *Business & Information Systems Engineering*, 66(1):111–126, 2024.
13 Carlos Fernández-Loría and Foster Provost. Causal classification: Treatment effect estimation vs. outcome prediction. *Journal of Machine Learning Research*, 23(59):1–35, 2022.
14 Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same… and why it matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022.
15 Carlos Fernández-Loría, Foster Provost, Jesse Anderton, Benjamin Carterette, and Praveen Chandar. A comparison of methods for treatment assignment with an application to playlist generation. *Information Systems Research*, 34(2):786–803, 2023.
16 Carlos Fernández-Loría, Foster Provost, and Xintian Han. Explaining data-driven decisions made by AI systems: The counterfactual approach. *Management Information Systems Quarterly*, 46(3):1635–1660, 2022.
17 Stefan Feuerriegel, Yash Raj Shrestha, Georg von Krogh, and Ce Zhang. Bringing artificial intelligence to business management. *Nature Machine Intelligence*, 4(7):611–613, 2022.
18 Travis Greene, Amit Dhurandhar, and Galit Shmueli. Atomist or holist? A diagnosis and vision for more productive interdisciplinary AI ethics dialogue. *Patterns*, 4(1):100652, 2023.
19 Sebastian Gabel, Daniel Guhl, and Daniel Klapper. P2v-map: Mapping market structures for large retail assortments. *Journal of Marketing Research*, 56(4):557–580, 2019.
20 Robin M Gubela, Stefan Lessmann, and Björn Stöcker. Multiple treatment modeling for target marketing campaigns: A large-scale benchmark study. *Information Systems Frontiers*, 26(3):875–898, 2024.
21 Travis Greene, David Martens, and Galit Shmueli. Barriers to academic data science research in the new realm of algorithmic behaviour modification by digital platforms. *Nature Machine Intelligence*, 4(4):323–330, 2022.

22 Travis Greene, Galit Shmueli, and Soumya Ray. Taking the person seriously: Ethically aware is research in the era of reinforcement learning-based personalization. *Journal of the Association for Information Systems*, 24(6):1527–1561, 2023.
23 Galit Shmueli, Ali Tafti. How to "improve" prediction using behavior modification, International Journal of Forecasting, Vol. 39(2), pp. 541–555, 2023.
24 Sebastian Gabel and Artem Timoshenko. Product choice with large assortments: A scalable deep-learning model. *Management Science*, 68(3):1808–1827, 2022.
25 Tobias Hatt and Stefan Feuerriegel. Early detection of user exits from clickstream data: A markov modulated marked point process model. In *Proceedings of The Web Conference 2020*, pages 1671–1681, 2020.
26 Tobias Hatt and Stefan Feuerriegel. Detecting user exits from online behavior: A duration-dependent latent state model. *arXiv preprint arXiv:2208.03937*, 2022.
27 Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, and Christian Janiesch. Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. *International Journal of Information Management*, 69:102538, 2023.
28 Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.
29 Milan Kuzmanovic, Dennis Frauen, Tobias Hatt, and Stefan Feuerriegel. Causal machine learning for cost-effective allocation of development aid. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.
30 Mathias Kraus, Stefan Feuerriegel, and Asil Oztekin. Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, 281(3):628–641, 2020.
31 Mathias Kraus, Stefan Feuerriegel, and Maytal Saar-Tsechansky. Data-driven allocation of preventive care with application to diabetes mellitus type II. *Manufacturing & Service Operations Management*, 26(1):137–153, 2023.
32 Abdurahman Maarouf, Stefan Feuerriegel, and Nicolas Pröllochs. A fused large language model for predicting startup success. *European Journal of Operational Research*, 2024.
33 David Martens and Foster Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–100, 2014.
34 Christof Naumzik, Stefan Feuerriegel, and Markus Weinmann. I will survive: Predicting business failures from customer ratings. *Marketing Science*, 41(1):188–207, 2022.
35 Yilmazcan Ozyurt, Tobias Hatt, Ce Zhang, and Stefan Feuerriegel. A deep Markov model for clickstream analytics in online shopping. In *ACM Web Conference*, 2022.
36 Yilmazcan Ozyurt, Mathias Kraus, Tobias Hatt, and Stefan Feuerriegel. AttDMM: An attentive deep Markov model for risk scoring in intensive care units. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.
37 Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc, 2013.
38 David Rügamer, Chris Kolb, and Nadja Klein. Semi-structured distributional regression. *The American Statistician*, 78(1):88–99, 2024.
39 Yash Raj Shrestha, Shiko M. Ben-Menahem, and Georg von Krogh. Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4):66–83, 2019.
40 Klaus Schwab. *The fourth industrial revolution.* World Economic Forum, 2017.
41 Galit Shmueli. Research dilemmas with behavioral big data. *Big Data*, 5(2):98–119, 2017.
42 Julian Senoner, Bernhard Kratzwald, Milan Kuzmanovic, Torbjørn H Netland, and Stefan Feuerriegel. Addressing distributional shifts in operations management: The case of order

fulfillment in customized production. *Production and Operations Management*, 32(10):3022–3042, 2023.

43 Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

44 Julian Senoner, Torbjørn Netland, and Stefan Feuerriegel. Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science*, 68(8):5704–5723, 2022.

45 Julian Senoner, Simon Schallmoser, Bernhard Kratzwald, Stefan Feuerriegel, and Torbjørn Netland. Explainable AI improves task performance in human-AI collaboration. *arXiv:2406.08271*, 2024.

46 Daniel Tschernutter and Stefan Feuerriegel. Data-driven dynamic police patrolling: An efficient Monte Carlo tree search. *European Journal of Operational Research*, 321(1):177–191, 2025.

Participants

- Kevin Bauer
Universität Mannheim, DE
- Margrét Bjarnadóttir
University of Maryland – College Park, US
- Jessica M. Clark
University of Maryland – College Park, US
- Theodoros Evgeniou
INSEAD – Fontainebleau, FR
- Carlos Fernández-Loria
HKUST – New Territories, HK
- Stefan Feuerriegel
Ludwig-Maximilians-Universität München, DE
- Sebastian Gabel
Erasmus University – Rotterdam, NL
- Travis Greene
Copenhagen Business School, DK
- Jungpil Hahn
National University of Singapore, SG
- Christian Janiesch
TU Dortmund, DE
- Enric Junqué de Fortuny
IESE Business School – Barcelona, ES
- Nadja Klein
KIT – Karlsruher Institut für Technologie, DE
- Mathias Kraus
Universität Erlangen-Nürnberg, DE
- Niklas Kühl
Universität Bayreuth, DE
- David Martens
University of Antwerp, BE
- Claudia Perlich
Two Sigma Investments LP – New York, US
- Joel Persson
Spotify – London, GB
- Foster Provost
New York University, US
- Galit Shmueli
National Tsing Hua University – Hsinchu, TW
- Sriram Somanchi
University of Notre Dame, US
- Wei Sun
IBM TJ Watson Research Center – Yorktown Heights, US
- Wouter Verbeke
KU Leuven, BE
- Michael Vössing
KIT – Karlsruher Institut für Technologie, DE
- Alona Zharova
HU Berlin, DE
- Patrick Zschech
Universität Leipzig, DE

Report from Dagstuhl Seminar 24351

Power, Energy, and Carbon-Aware Computing on Heterogeneous Systems (PEACHES)

Kerstin I. Eder*[1], Timo Hönig*[2], Maja Hanne Kirkeby*[3], Daniel Mosse*[4], Max Plauth*[5], and Jonas Juffinger†[6]

1 **University of Bristol, GB.** `cskie@bristol.ac.uk`
2 **Ruhr-Universität Bochum, DE.** `timo.hoenig@rub.de`
3 **Roskilde University, DK.** `majaht@ruc.dk`
4 **University of Pittsburgh, US.** `mosse@cs.pitt.edu`
5 **UltiHash – Berlin, DE.** `max.plauth@icloud.com`
6 **TU Graz, AT.** `jonas.juffinger@iaik.tugraz.at`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 24351 "Power, Energy, and Carbon-Aware Computing on Heterogeneous Systems (PEACHES)".

Seminar August 25–30, 2024 – https://www.dagstuhl.de/24351

2012 ACM Subject Classification Computer systems organization → Heterogeneous (hybrid) systems; General and reference; General and reference → Evaluation; General and reference → Measurement; General and reference → Metrics; Hardware → Power and energy; Software and its engineering → Operating systems

Keywords and phrases Carbon, Energy, Operating Systems, Power, Ressource Efficiency, Systems

Digital Object Identifier 10.4230/DagRep.14.8.36

1 Executive Summary

Kerstin I. Eder (University of Bristol, GB)
Timo Hönig (Ruhr-Universität Bochum, DE)
Maja Hanne Kirkeby (Roskilde University, DK)
Daniel Mosse (University of Pittsburgh, US)
Max Plauth (UltiHash – Berlin, DE)

The increasing carbon footprint of computing systems represents a critical and immediate challenge for the computing community, particularly as global digitalization accelerates. As computational demands grow – driven by data-intensive applications such as artificial intelligence – addressing the environmental impact of computing has become essential. The Dagstuhl Seminar 24351 on Power, Energy, and Carbon-Aware Computing on Heterogeneous Systems (PEACHES) focused on advancing the field's understanding of how carbon emissions can be effectively measured, managed, and reduced across all layers of heterogeneous computing systems.

This seminar (PEACHES) brought together leading researchers and practitioners from the fields of computer science, software engineering, and environmental sustainability. The seminar participants explored 5 essential topics through continued group discussions. The

* Editor / Organizer
† Editorial Assistant / Collector

participants explored boundaries, challenges, and possible new methods and techniques for: 1) obtaining **Carbon Transparency**, 2) reaching **Net-Zero in the Age of AI and Machine Learning**, and 3) achieving **Carbon-Aware Computing, Storage, and Communication**. In addition, we discussed and explored **Disruptive Paradigms**, focusing on innovative approaches needed to achieve net-zero carbon goals, rather than relying on incremental improvements. The seminar also featured enlightening **Carbon-aware Computing Hackathons**, introducing and discussing the latest research tools in the area, e.g., to reduce energy consumption via undervolting and tracking carbon emissions in applications. The seminar contributes to advancing the understanding of how software innovations can help mitigate the global environmental impact of computing.

2 Table of Contents

Working groups

3 Overview of Talks

3.1 Performance is not efficiency

Gustavo Alonso (ETH Zürich, CH)

Performance measures how well a system completes a task in terms of throughput and latency, typically with the aim of maximizing the first and minimizing the latter. Efficiency measures how well the resources needed to accomplish a task are used focusing on reducing, e.g., utilization and idle times. The IT industry is focused almost exclusively on performance as seen in the common service level agreements used in industry and in benchmarks. Efficiency plays a role mainly in terms of overall monetary costs and does not necessarily include environmental impacts or externalized costs. In the talk I will discuss the difference between performance and efficiency in relation to Green IT with the goal of identifying the biggest targets for potential savings.

3.2 Tape-based storage reduces CO2 emissions

Pierre Bennorth (CopenCloud – Karlslunde, DK)

In today's digital age, vast amounts of data are stored on servers, much of which is rarely accessed but continuously consumes energy. At CopenCloud, we have developed a solution that categorizes data into two distinct groups: frequently accessed data stored on hard drives for instant access, and infrequently used data stored on energy-efficient tape drives. This approach allows us to significantly reduce energy consumption and CO2 emissions, with up to 90% savings compared to traditional storage methods.

Our CO2 reduction claims have been validated by the esteemed Danish consultancy Rambøll, whose report [1] confirms that our tape-based storage reduces CO2 emissions by 90% compared to hard drives. The report also highlights that globally, 5.6 zettabytes of data are ideal for long-term tape storage, and around 80% of organizational data can benefit from this method.

Additionally, for every 17 terabytes stored on tape, we can save one ton of CO2 annually. To aid organizations in optimizing their data storage, we've developed a Directory Scanner that identifies data suitable for long-term storage. We also provide a CO2 report for ESG accounts, ensuring our clients can track their environmental impact.

Our solution prioritizes security, featuring write protection via S3 and Glacier Protocol, making it resilient against ransomware attacks. Furthermore, our platform is designed to minimize energy consumption while enabling faster uploads and downloads. We are collaborating with Maja H. Kirkeby on the project "Software's energy consumption at the end-user device" to have these capabilities independently verified.

References

1 Rambøll, "*CopenCloud Rapport Intelligent data storage can optimize organizations' costs and lower their climate footprint.*" Available at: `https://copencloud.com/docs/CopenCloud_Report_EN.pdf`, 2023. Accessed: 2024-09-12.

3.3 Heterogeneous Computing for Energy with the TeamPlay Coordination Language

Clemens Grelck (Friedrich-Schiller-Universität Jena, DE)

Joint work of Clemens Grelck, Sebastian Altmeyer, Benjamin Rouxel, Julius Roeder, Andy D. Pimentel, Lukas Miedema, Ulrik Pagh Schultz, Jesper Holst, Ole Jørgensen, Benny Åkesson
Main reference Julius Roeder, Benjamin Rouxel, Sebastian Altmeyer, Clemens Grelck: "Towards Energy-, Time- and Security-Aware Multi-core Coordination", in Proc. of the Coordination Models and Languages – 22nd IFIP WG 6.1 International Conference, COORDINATION 2020, Held as Part of the 15th International Federated Conference on Distributed Computing Techniques, DisCoTec 2020, Valletta, Malta, June 15-19, 2020, Proceedings, Lecture Notes in Computer Science, Vol. 12134, pp. 57–74, Springer, 2020.
URL https://doi.org/10.1007/978-3-030-50029-0_4

Heterogeneous compute architectures are key to energy savings, provided that software chooses to components of software systems on the individually best suited hardware components available. High-performance embedded devices and other battery-powered devices from mobile phones to laptops are typically equipped with various instances of multiple types of CPU and GPU cores with highly different performance/energy trade-offs and characteristics. The open question is how applications can be mapped to such architectures under global objectives such as meeting deadlines or achieving required performance levels while using the minimal amount of energy possible.

The TeamPlay coordination language is geared towards describing componentised applications, their dependencies, constraints, and objectives on a high level of abstraction. As a coordination language TeamPlay delegates the implementation of software components of reasonable size and complexity to established programming languages for programming-in-the-small and entirely focusses on coordination aspects such as time-wise and spatial scheduling, dependencies and communication, etc, all in line with the aforementioned global objectives. Compiler and runtime system middleware are charged with establishing transparency about the inevitable trade-offs and actually run applications on the heterogeneously parallel system architectures targeted.

References

1 J. Roeder, B. Rouxel, S. Altmeyer, C/ Grelck. *Towards Energy-, Time- and Security-aware Multi-core Coordination*. 22nd International Conference on Coordination Models and Languages (COORDINATION 2020), virtual from Valletta, Malta, Lecture Notes in Computer Science 12134, pp. 57-74.

2 B. Rouxel, U. Pagh Schultz, B. Åkesson, J. Holst, O. Jørgensen, C. Grelck. *PReGO: a generative methodology for satisfying real-time requirements on COTS-based systems: Definition and experience report*. 19th ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences (GPCE 2020), virtual from Chicago, USA, ACM, 2020.

3 J. Roeder, B. Rouxel, S. Altmeyer, C. Grelck. *Energy-aware Scheduling of Multi-version Tasks on Heterogeneous Real-time Systems*. 36th ACM/SIGAPP Symposium on Applied Computing (SAC 2021), pp. 500–510, ACM, 2021.

4 B. Rouxel, S. Altmeyer, C. Grelck. *YASMIN: a Real-time Middleware for COTS Heterogeneous Platforms*. 22nd ACM/IFIP International Middleware Conference (MIDDLEWARE 2021), pp. 298-309, virtual from Montréal, Canada, ACM, 2021.

5 J. Roeder, A.D. Pimentel, C. Grelck. *GCN-based Reinforcement Learning Approach for Scheduling DAG Applications*. Artificial Intelligence Applications and Innovations, 19th IFIP WG 12.5 International Conference (AIAI 2023), León, Spain, IFIPAICT Vol. 676, pp. 121-134, Springer, 2023.

3.4 Optimizing Green Coding Practices: Measurement Accuracy and Best Practices

Geerd-Dietger Hoffmann (Green Coding Solution – Berlin, DE)

In the quest for sustainable and efficient software, accurate energy consumption measurement is paramount. This talk addresses common pitfalls and best practices in green coding, specifically focusing on the importance of understanding the tools and techniques used for energy measurement in software development. Through real-world examples and critical analysis, we explore the limitations of popular measurement tools like Intel RAPL and the Gude Power Meter, and discuss the necessity of consistent temperature conditions, disabling CPU features like Turbo Boost and Hyper-Threading, and the challenges of using CPU utilization as a reliable metric. Moreover, we emphasize the significance of validating machines for anomalies, automating processes, and considering the overhead of measurement setups. Practical advice on using specialized operating systems, non-standard sampling intervals, and handling external network traffic is also provided. This talk invites discussion and collaboration on improving green coding practices, ultimately aiming to refine how we measure and reduce the environmental impact of software development.

3.5 Quantifying the potential of link sleeping in ISP networks

Romain Jacob (ETH Zürich, CH)

Main reference Lukas Röllin, Romain Jacob, Laurent Vanbever: "A Sleep Study for ISP Networks: Evaluating Link Sleeping on Real World Data", 2024.
URL https://doi.org/10.3929/ethz-b-000680859

It is widely accepted that energy proportionality, i.e., the ability to scale the power drawn by a system with its workload, is an enabler for energy efficiency. Putting parts of a system to sleep – turning them off or into some low-power mode – when they are not needed is the B-A-BA of proportionality. In networks, a simple part one can turn off is a link: if traffic volume is low, one can often route traffic away from some links that can be put to sleep to save energy. But how to select the links one can safely turn off without causing congestion elsewhere in the network? And how much energy would that actually save?

Our ongoing research addresses these questions. We found that, in lowly loaded networks, simple heuristics appear sufficient to decide which links to turn off. Using real topology and traffic data from two small ISPs (Switch (CH) and SURF (NL)), we found that about 1/3 of links can be turned off on average. As I will discuss in the talk, the conversion into expected energy savings requires further research.

3.6 Cloud Server Design for Lower Carbon

Fiodar Kazhamiaka (Microsoft – Redmond, US)

Joint work of Fiodar Kazhamiaka, Jaylen Wang, Daniel Berger, Fiodar Kazhamiaka, Celine Irvene, Chaojie Zhang, Esha Choukse, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, Ricardo Bianchini, Akshitha Sriraman
Main reference Jaylen Wang, Daniel S. Berger, Fiodar Kazhamiaka, Celine Irvene, Chaojie Zhang, Esha Choukse, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, Ricardo Bianchini, Akshitha Sriraman: "Designing Cloud Servers for Lower Carbon", in Proc. of the 51st ACM/IEEE Annual International Symposium on Computer Architecture, ISCA 2024, Buenos Aires, Argentina, June 29 – July 3, 2024, pp. 452–470, IEEE, 2024.
URL https://doi.org/10.1109/ISCA59077.2024.00041

Cloud servers can be designed to optimize for carbon emissions. Hardware developments such as CXL and dense/efficient CPU cores offer improved performance/Watt and the ability to give a second life to older components with high embodied carbon, such as DIMMs. The talk covers the design of such servers, and a framework for evaluating the carbon savings at the datacenter, including features such as application performance, customer adoption, maintenance implications, and workload packing implications.

3.7 Energy Labelling in Software

Maja Hanne Kirkeby (Roskilde University, DK)

Joint work of Maja Hanne Kirkeby, Kerstin I. Eder, E. B. Unna-Lindhard, N. Müllenborn

This presentation explores the energy consumption associated with Information and Communication Technology (ICT) systems, with a particular focus on how software influences overall energy use. It proposes the introduction of energy labels for software and software modules. ICT's contribution to global greenhouse gas (GHG) emissions is substantial, accounting for an estimated 1.8 % to 2.8 % of total emissions, primarily driven by the operational phases of these systems [2]; e.g., the GHG emissions was evaluated to be 1.4 % in 2020 for the operational phase [1]. Existing research demonstrates that optimization at the software level can lead to significant energy savings. For example, replacing specific JavaScript libraries in web applications can reduce energy consumption by up to 30 % [3]. Despite the potential for considerable energy savings, a large portion of software developers do not prioritize energy efficiency during development, with only about 10% attempting to measure energy consumption [4]. We propose the creation of energy labels for software, inspired by the EU energy labels for household products. These labels are intended to provide relevant, comparable, and accurate information on energy efficiency, thereby enabling users and developers to make informed decisions. The EU labels have been recognized as "one of the most cost-effective ways to enhance security of energy supply..." [5]. Our preliminary study on WordPress plugins – software modules – indicates that (1) it is feasible to assess individual software modules independently when investigating the energy consumption of web-based software, (2) it is possible to measure and determine their impact on the energy consumption of the entire system, and (3) this approach allows for the first comparison of energy consumption among individual modules offering similar functionalities [6]. The study suggests that it is feasible to develop energy labels for software in the future, providing relevant, comparable, and accurate information on energy efficiency. We envision creating energy labels for functionally equivalent modules and, in the long run, standardizing the process for creating such labels to be applicable across different levels within the software stack, effectively driving reductions in the environmental impact of software.

References

1 Jens Malmodin, Nina Lövehagen, Pernilla Bergmark, Dag Lundén, "ICT sector electricity consumption and greenhouse gas emissions – 2020 outcome", Telecommunications Policy, Volume 48, Issue 3, 2024, https://doi.org/10.1016/j.telpol.2023.102701.
2 C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. Blair, and A. Friday, "The climate impact of ICT: A review of estimates, trends, and regulations," *December*, 2020.
3 A. P. Miettinen and J. K. Nurminen, "Analysis of the Energy Consumption of JavaScript Based Mobile Web Applications," in *International Conference on Green IT*, 2010, doi: 10.1007/978-3-642-16644-0_12.
4 C. Pang, A. Hindle, B. Adams, and A. E. Hassan, "What do programmers know about the energy consumption of software?," *PeerJ PrePrints*, vol. 3, p. e1094, 2015.
5 Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, "Energy Efficiency Plan 2011," *Issue 1*, 2011.
6 E. B. Unna-Lindhard and N. Müllenborn, "Energy consumption of WordPress plugins," 2022.

3.8 10 years of GCC from an energy perspective

Michael Kirkedal Thomsen (University of Copenhagen, DK), Maja Hanne Kirkeby (Roskilde University, DK)

Joint work of Michael Kirkedal Thomsen, Ken Friis Larsen, Maja Hanne Kirkeby, Lars-Bo Husted Vadgaard

In this work, we present an investigation into 10 years of GCC versions from an energy perspective. Specifically, we employ Intel's Running Average Power Limit (RAPL) technology to compare the energy consumption of different versions of GCC when they compile a relevant subset of the SPEC benchmark. Initial results show no significant improvement and may suggest that older GCC versions are more efficient at compiling. However, more measurements are needed to obtain a clear statistical result.

3.9 Trace & Tag: Towards Carbon-Aware Memory Placement

Sven Köhler (Hasso-Plattner-Institut, Universität Potsdam, DE) and Timo Hönig (Ruhr-Universität Bochum, DE)

Joint work of Sven Köhler, Wenzel Lukas, Wenzel, Timo Hönig, Andreas Polze

Memory subsystems have long been overlooked consumers of operational energy in computing systems. Novel memory technologies, like NVRAM, further introduce wear-and-tear behaviour, making embodied carbon a factor in memory placement decisions.

With cMemento we aim to supply system software with knowledge about workload behaviour and hardware resources alike for more carbon-aware steering of allocations.

One key requirement is an understanding of a workload's sub-page memory access patterns. Hence, we propose Heimdall, a novel memory device that occupies part of the physical address pace. It is backed by other DRAM in the system but can snoop on and redirect requests on the memory bus. Our prototype runs on an AMD Zynq 7000 SOC, with an ARM Cortex-A and tightly integrated Xilinx FPGA.

Heimdallr can simulate heterogenous memory devices, using the FPGA's block RAM. We present an architecture and operating system interface to trace and process access patterns on the memory bus, without impeding the overall bandwidth. Although we add about 90% latency while the tracing is enabled, an early evaluation shows that we can use the gained knowledge to move the most promising allocations to the block RAM – saving an average of 11% in operational energy permanently.

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Individual Research Grant 502228341 (Memento) as part of the Priority Programme on Disruptive Memory Technologies (SPP 2377).

3.10 Working towards a standardized reporting and assessment of the energy efficiency of ML models training and inference

Silverio Martínez-Fernández (UPC Barcelona Tech, ES)

Main reference Silverio Martínez-Fernández, Xavier Franch, Francisco Durán: Towards green AI-based software systems: an architecture-centric approach (GAISSA). SEAA 2023: 432-439
URL https://doi.org/10.1109/SEAA60479.2023.00071

Software-related CO2 emissions from the information and communications technology sector currently account for 2.1%–3.9% of global emissions. With the latest advancements in Machine Learning (ML) systems, this percentage of global emissions is estimated to increase. This raises the question: how to create more awareness on the sustainability impact of training and inference of ML models? In this talk, I navigate through the capabilities of the GAISSALabel tool [1] form the GAISSA project [2]. The tool shows how to report green ML metrics, to measure and assess the energy efficiency in ML systems, and shows recommendations on how to reduce their carbon emissions (e.g., reducing ML model size, using lightweight architectures, early stopping, optimization tactics as quantization and pruning, energy-efficient ML serving, and when to deploy in cloud or edge). Efforts to standardize the reporting of ML carbon emissions are needed (e.g., ISO 20226, SCI from GSF). Besides reporting, the aggregation of the results of empirical studies from the green software engineering community, in order to create theories with consolidated and quantified software tactics to reduce emissions, is key.

References

1 Pau Duran, Joel Castaño, Cristina Gómez, Silverio Martínez-Fernández: GAISSALabel: A Tool for Energy Labeling of ML Models. SIGSOFT FSE Companion 2024: 622-626
2 Silverio Martínez-Fernández, Xavier Franch, Francisco Durán: Towards green AI-based software systems: an architecture-centric approach (GAISSA). SEAA 2023: 432-439

3.11 CXL Memory Expansion: A Closer Look on Actual Platforms

Vinicius Petrucci (Micron Technology – Austin, US)

The proliferation of data-intensive workloads has led to a significant increase in the volume of data that computing systems need to handle. The growing significance of memory as a bottleneck has resulted in increased total cost of ownership (TCO) for servers. Existing methods to enhance bandwidth involve raising the data rate; while scaling up capacity is achieved through techniques like 2DPC or increasing DRAM density, each of which are coming at the expense of the other. An alternative approach is to scale out by adding more CPUs or through 3D stacking memory packaging technology, but these approaches come at the expense of much higher TCO.

In this talk I will discuss the potential use cases of Compute Express Link (CXL) technology. CXL is an open industry standard interconnect offering coherency and memory semantics using high bandwidth, low latency connectivity between processors, accelerators, memory, storage, and other IO devices. It helps address the above challenges by scaling up through increased bandwidth and capacity per core in the system. CXL is gaining faster adoption because it introduces load/store semantics to PCI Express (PCIe) physical layer, enabling expansion of both capacity and bandwidth at access latency comparable to remote non uniform memory access (NUMA) node DRAM memory. This addresses the CPU pin count challenge and avoids the TCO associated with adding new servers and processors just to get additional bandwidth and capacity.

3.12 Reducing Storage Emissions based on Deduplication

Max Plauth (UltiHash – Berlin, DE)

URL https://ultihash.io/whitepaper

Until recently, the emissions caused by storage resources were largely disregarded, with both academic and industry efforts having largely focused on emissions attributed to compute resources. To keep up with the excessive demand for storage resources across disciplines and industries however, all available opportunities need to be explored to enable data volumes to grow more sustainably. While research efforts on reducing storage emissions are still in an early stage, they largely focus on three general strategies to reduce storage emissions: By lowering and/or shifting power, by increasing the density of storage devices, or by extending the lifetime of storage hardware [1].

With the development of the UltiHash object storage, we are proposing an additional strategy to reduce storage emissions. We propose that using fine-grained deduplication, we can reduce the number of storage devices required to store data. With this approach, we are accepting slightly increased operational emissions caused by spending CPU time once to reduce both operational and embodied emissions of storage devices over the retention time of the stored data.

Our approach differs from existing deduplication mechanisms in two distinctive ways: first, using a variable fragment size enables us to detect more fine grained occurrences of redundancies. Second, instead of operating on a limited scope such as s single file system or a single file server, the scope of deduplication spans across a distributed storage cluster, aiming even towards federations of storage clusters.

While deduplication rates of our approach highly depend on the properties of the dataset to be deduplicated, initial tests have yielded data deduplication rates of up to 70% for uncompressed datasets, and even deduplication rates >= 10% for compressed media files.

References

1 Sara McAllister, Fiodar Kazhamiaka , Daniel S. Berger, Rodrigo Fonseca, Kali Frost, Aaron Ogus, Maneesh Sah, Ricardo Bianchini, George Amvrosiadis, Nathan Beckman, Gregory R. Ganger, "A call for research on storage emissions", *3rd Workshop on Sustainable Computer Systems (HotCarbon)*, 2024

3.13 The Cost of a Secure OS

Fabian Rauscher (TU Graz, AT), Daniel Gruss (TU Graz, AT), and Timo Hönig (Ruhr-Universität Bochum, DE)

In the security community it is commonly assumed that performance and energy are correlated and it is sufficient to just measure performance overheads to assess the overall costs of a security measure. We investigated whether this is true for the Linux kernel and found that in a significant number of cases energy and performance costs are actually not correlated. We conclude that to assess the energy costs of security, it is currently unavoidable to perform actual energy measurements.

3.14 Green AI: The more does not (always) lead to the better!

June Sallou (TU Delft, NL)

Joint work of Nienke Nijkamp, June Sallou, Niels van der Heijden, Luís Cruz
Main reference Nienke Nijkamp, June Sallou, Niels van der Heijden, Luís Cruz: "Green AI in Action: Strategic Model Selection for Ensembles in Production", in Proc. of the 1st ACM International Conference on AI-Powered Software, AIware 2024, Porto de Galinhas, Brazil, July 15-16, 2024, ACM, 2024.
URL https://doi.org/10.1145/3664646.3664763

The integration of Artificial Intelligence (AI) into software systems has significantly enhanced their functional capabilities, albeit at the cost of increased energy demands. Ensemble learning, which combines predictions from multiple models to derive a single output, exacerbates this issue by intensifying cumulative energy consumption. We address these trade-off challenges through an industrial case study of an AI-enabled system using ensemble learning for document information extraction. Our findings demonstrate that our approach of selecting a subset of models by optimising the GreenQuotientIndex, which accounts for the trade-off between accuracy and resource usage, can lead to resource savings of up to 99 %. Additionally, this approach reveals unexpected gaps in system optimisation related to accuracy. Selecting a subset of AI models can even increase accuracy by up to a factor of 2.7. These insights underscore the necessity of defining sustainability as a critical requirement in software development, rather than defaulting to a higher number of models without careful consideration.

3.15 Executable explanations of control software for wind turbines

Sibylle Schupp (TU Hamburg, DE)

Cyber-physical energy systems (CPES) are digitally controlled systems in which the degree of heterogeneity and complexity has become so high that their particular behavior is no longer necessarily transparent – not even to developers or users with technical background. One way of providing more transparency is by giving an explanation of the particular behavior of interest; and for that, the subsystem in question is in the best positions, as it best could explain itself. So that a self-explanation is operationizable by its addressee, it needs to be geared towards that specific addressee. Also, it has to be provided timely and in a context-sensitive fashion, and in particular at a level of abstraction different from the behavior that should be explained.

Consider as example a wind turbine controller. Characteristic for controllers is the feedback loop, in which the input depends non-trivially on the output as well as the interaction with other control systems, e.g., the controller for the collective pitch angle interacts with the controller for the yaw. An addressee of an explanation of that subsystem might be a maintainer or an optimizer or fault searcher as well as an operator of a client system. Now assume the addressee observes an unexpected output value and thus is in need of an explanation. So that the addressee can properly respond, it must not consider one such parameter just in isolation. On the other hand, it cannot be expected to understand the full control-theoretic details of its dependences. Instead, the addressee needs a simplified, abstract explanation of the system behavior, here: of the turbine controller.

For reducing the complexity of control software while keeping mutual dependencies of the feedback loop, we exploit that the differential equations of the control problem can be piecewise approximated by simple (linear) functions and that in an online setting those simplifying "control patterns" can be swapped out when outdated. The overall soundness of the approximation then depends on the "control patterns", the choice of time windows, and the safe and timely swapping of patterns.

The talk motivates the abstract explanation pattern introduced by the CAUSE (Concepts and Algorithms for–and Usage of–Self-Explaining Systems) Research Training Group and, for wind turbines, illustrates this explanation pattern on its control software where explanations are executable networks of timed automata.

3.16 How to quantify the energy consumption of applications

Tamar Eilam (BM TJ Watson Research Center – Yorktown Heights, US)

Improving the efficiency of software means running more computing using less resources, and energy. However, in order to achieve this goal we need to first be able to quantify the energy consumption of applications. Quantification is hard for multiple reasons: in realistic environments multiple applications run on shared infrastructure. First, there is no way to measure the energy for individual applications, hence we need to approximate it. Second, in many cases there is no access to the actual on-line power reads of the machine (e.g., in cloud environment). Third, applications behave differently on different hardware and configuration leading to combinatorial explosion. In this talk I will describe the approach we took with our open source tool Kepler to address these challenges.

References

1 https://sustainable-computing.io

3.17 Energy-efficient systems through non-volatile memory

Manuel Vögele (Ruhr-Universität Bochum, DE), Timo Hönig (Ruhr-Universität Bochum, DE), and Wolfgang Schröder-Preikschat (Universität Erlangen-Nürnberg, DE)

Main reference Jonas Rabenstein, Dustin T. Nguyen, Oliver Giersch, Christian Eichler, Timo Hönig, Jörg Nolte, Wolfgang Schröder-Preikschat: "Back to the Core-Memory Age: Running Operating Systems in NVRAM only", in Proc. of the Architecture of Computing Systems – 36th International Conference, ARCS 2023, Athens, Greece, June 13-15, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 13949, pp. 153–167, Springer, 2023.
URL https://doi.org/10.1007/978-3-031-42785-5_11

Non-volatile RAM (NVRAM) is fast, byte-addressable, non-volatile memory that can replace traditional DRAM in systems. Doing so improves the non-functional properties of a system, including lower energy usage, faster storage and higher resiliency. This talk focuses on the energy usage aspect.

There are several opportunities to save energy by using NVRAM. First, opposed to DRAM, NVRAM does not need to be refreshed, leading to low idle power consumption. Second, by getting rid of all volatile memory in our systems, they become able to rapidly power off and resume their operation. This can be used to shut off machines during even short periods of idle time, while staying able to quickly respond to requests. Third, a lot of persistency mechanisms in operating systems are no longer necessary, as NVRAM provides that persistence for free. This means we can declutter our operating systems, getting rid of energy intensive I/O operations in the process. However, there are still challenges to overcome to realize those savings. Today's systems are built with DRAM in mind. Booting into NVRAM is challenging and slow firmware hinders rapid suspend/resume mechanisms. Finally, introducing a new class of memory into a system requires the operating system to perform placement decisions to ensure the heterogeneous memory resources are used most efficiently.

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Individual Research Grant 465958100 (NEON).

4 Working groups

4.1 Hackathon: CPU Undervolting

Jonas Juffinger (TU Graz, AT)

Joint work of Jonas Juffinger, Stepan Kalinin, Daniel Gruss, Frank Mueller
Main reference Jonas Juffinger, Stepan Kalinin, Daniel Gruss, Frank Mueller: "SUIT: Secure Undervolting with Instruction Traps", in Proc. of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024, pp. 1128–1145, ACM, 2024.
URL https://doi.org/10.1145/3620665.3640373

CPU undervolting, running a CPU with a lower voltage than specified by the manufacturer, can save up to 20% of CPU power consumption. However, CPU undervolting also negatively impacts systems' reliability and security, which is not tolerable in most use cases. In our work, SUIT, we showed how CPU undervolting could be made secure to allow the usage of it on a wide scale. In this hackathon, we prepared some experiments to measure the energy and performance impact of CPU undervolting. We measured up to 20 % reduction in energy consumption without any performance impact. Another experiment could be used to play with the undervolting offset and experience the possible reliability issues. Finally, we prepared experiments to show the possible security issues of CPU undervolting and how it can be used steal cryptographic keys. About 30 people participated in the hackathon.

4.2 Hackathon: Carbond

Robin Ohs (Universität des Saarlandes – Saarbrücken, DE), Benedict Herzog, Gregory Stock, Henry Janson, Timo Hönig (Ruhr-Universität Bochum, DE), Luis Gerhorst, and Andreas Schmidt (Universität des Saarlandes – Saarbrücken, DE)

Main reference Andreas Schmidt, Gregory Stock, Robin Ohs, Luis Gerhorst, Benedict Herzog, Timo Hönig: "carbond: An Operating-System Daemon for Carbon Awareness", in Proc. of the 2nd Workshop on Sustainable Computer Systems, HotCarbon 2023, Boston, MA, USA, 9 July 2023, pp. 2:1–2:6, ACM, 2023.
URL https://doi.org/10.1145/3604930.3605707

The carbon intensity of the grid is influenced by the mix of carbon emissions generated by various electricity production methods. To utilize this intensity metric for carbon-aware computing, a common approach is to query data providers such as "Electricity Maps" or "WattTime" to obtain the average grid intensity of the desired location. Although this is an effective initial approach, it underestimates the actual carbon intensity of the utilized energy due to inefficiencies within the system itself coming from the power supply unit (AC to DC conversion) and, if battery powered, from the storage of the energy. The presented tool Carbond is an operating system daemon that aims to calculate a system intensity that approximates more accurately the intensity of the energy you are currently using. The system intensity is calculated by considering the above-mentioned inefficiencies and an estimated embodied intensity of the battery. This hackathon explains the principles and origins of these inefficiencies and showcases the practical use of Carbond. For this, the participants measure the required energy of an energy-intensive workload and then connect this energy demand to actual emissions.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project numbers 502228341 ("Memento"), 465958100 ("NEON"), and 389792660 as part of TRR 248 – CPEC4.

Participants

- Gustavo Alonso
ETH Zürich, CH
- Antonio Carlos Schneider Beck Filho
Federal University of Rio Grande do Sul, BR
- Pierre Bennorth
CopenCloud – Karlslunde, DK
- Liliana Cucu-Grosjean
INRIA – Paris, FR & StatInf – Paris, FR
- Ada Diaconescu
Telecom Paris, FR
- Kerstin I. Eder
University of Bristol, GB
- Tamar Eilam
IBM TJ Watson Research Center – Yorktown Heights, US
- Michael Engel
Universität Bamberg, DE
- João Paulo Fernandes
New York University – Abu Dhabi, AE
- Clemens Grelck
Friedrich-Schiller-Universität Jena, DE
- Daniel Gruss
TU Graz, AT
- Timo Hönig
Ruhr-Universität Bochum, DE
- Geerd-Dietger Hoffmann
Green Coding Solution – Berlin, DE
- Romain Jacob
ETH Zürich, CH
- Jonas Juffinger
TU Graz, AT
- Fiodar Kazhamiaka
Microsoft – Redmond, US
- Maja Hanne Kirkeby
Roskilde University, DK
- Michael Kirkedal Thomsen
University of Copenhagen, DK
- Sven Köhler
Hasso-Plattner-Institut, Universität Potsdam, DE
- David Kohnstamm
Leafcloud – Amsterdam, NL
- Julia Lawall
INRIA – Paris, FR
- Silverio Martínez-Fernández
UPC Barcelona Tech, ES
- Daniel Mosse
University of Pittsburgh, US
- Frank Mueller
North Carolina State University – Raleigh, US
- Jukka K. Nurminen
University of Helsinki, FI
- Robin Ohs
Universität des Saarlandes – Saarbrücken, DE
- Vinicius Petrucci
Micron Technology – Austin, US
- Max Plauth
UltiHash – Berlin, DE
- Fabian Rauscher
TU Graz, AT
- June Sallou
TU Delft, NL
- Joao Saraiva
University of Minho, PT
- Andreas Schmidt
Universität des Saarlandes – Saarbrücken, DE
- Gunnar Schomaker
Universität Paderborn, DE
- Wolfgang Schröder-Preikschat
Universität Erlangen-Nürnberg, DE
- Sibylle Schupp
TU Hamburg, DE
- Jennifer Switzer
University of California – San Diego, US
- Manuel Vögele
Ruhr-Universität Bochum, DE
- Samuel Xavier-de-Souza
Federal University of Rio Grande do Norte, BR

Report from Dagstuhl Perspectives Workshop 24352

Conversational Agents: A Framework for Evaluation (CAFE)

Christine Bauer*[1], Li Chen*[2], Nicola Ferro*[3], and Norbert Fuhr*[4]

1 **Paris Lodron University Salzburg, AT.** christine.bauer@plus.ac.at
2 **Hong Kong Baptist University, HK.** lichen@comp.hkbu.edu.hk
3 **University of Padua, IT.** nicola.ferro@unipd.it
4 **Universität Duisburg-Essen, DE.** norbert.fuhr@uni-due.de

Abstract

This report documents the program and the outcomes of the Dagstuhl Perspectives Workshop 24352, "Conversational Agents: A Framework for Evaluation (CAFE)", which brought together 22 distinguished researchers and practitioners from 12 countries. In this workshop, a new framework for the evaluation of conversational information access systems was developed, consisting of six major components: 1) goals of the system's stakeholders, 2) user tasks to be studied in the evaluation, 3) aspects of the users carrying out the tasks, 4) evaluation criteria to be considered, 5) evaluation methodology to be applied, and 6) measures for the quantitative criteria chosen. An evaluation design begins with identifying the stakeholders, whose goals determine the criteria. Tasks and evaluation methodology should be chosen according to these decisions.

Seminar August 25–30, 2024 – https://www.dagstuhl.de/24352
2012 ACM Subject Classification Information systems → Information retrieval; Information systems → Recommender systems; Computing methodologies → Natural language processing
Keywords and phrases Conversational Agents, Evaluation, Information Access, Dagstuhl Perspectives Workshop
Digital Object Identifier 10.4230/DagRep.14.8.53

1 Executive Summary

Christine Bauer
Li Chen
Nicola Ferro
Norbert Fuhr

In this Dagstuhl Perspectives Workshop, a general model for the evaluation of CONversational Information ACcess (CONIAC) systems was developed: Conversational Agents Framework for Evaluation (CAFE).

The framework starts from the assumption that a CONIAC system will be able to (i) *interact* with users more naturally and seamlessly, (ii) guide a user through the process of *refining* and *clarifying* their needs, (iii) *aid decision-making* by making *personalized recommendations and information* while being able to *explain* them, and (iv) *generate*, *retrieve* and *summarize* relevant information.

* Editor / Organizer

CAFE distinguishes six major elements of an evaluation design:

- **Stakeholder goals.** Stakeholders of a CONIAC system may have diverse goals that might or might not be directly accessible to system designers or evaluators and must often be implicitly inferred in evaluation. CONIAC systems might also have multiple goals ranging from end users having (in-)direct information needs, to platforms deploying CONIAC systems interested in content usage, user engagement, impression generation, and user retention, to name a few.
- **Tasks.** CONIAC involves tasks characterized by an information need (which may be specific or rather vague), human involvement, goal orientation, and mixed initiative between the user and the system. While some tasks and information needs may benefit from introducing a conversationally competent system, others may not, depending on the complexity of the task or need.
- **User aspects.** When developing an evaluation framework for CONIAC systems, it is crucial to consider user-specific aspects, such as preferences, specialized needs, expertise types, and background characteristics, which may make conversational systems more beneficial than non-conversational alternatives.
- **Criteria.** The scope of evaluation can range from single-turn interactions to entire conversations and long-term system usage, each requiring different criteria for assessment. Additionally, the temporal dimension, which examines how the system's performance changes over time, is a critical factor that can intersect with both stationary and dynamic properties. Criteria may be system-centric, user-centric, or both. The former regard hardware and software aspects like e. g. efficiency, accuracy, comprehensiveness, and verifiability. For the latter, we can distinguish between conversation-oriented (like e. g. adaptability, coherence, fluency), content-oriented (like e. g. continuance, controllability, perceived accuracy, understandability), and consequences-oriented measures (like e. g. addiction, benevolence, decision quality, confidence, trust).
- **Methodology.** In addition to the standard distinction of user-focused and system-focused methodologies, our evaluation framework categorizes evaluation methodologies also according to the employed time model – a dimension especially relevant for CONIAC. This dimension ranges from stationary methodologies like single-interaction experiments to methodologies like controlled lab studies that allow for continuous measurements such as physiological ones.
- **Measures.** Finally, we allow for measures that typically focus on the system's ability to provide accurate, relevant, and timely information during interactions. Measures include objective measures of effectiveness and subjective notions such as perceived effectiveness or user satisfaction (e. g., self-reported satisfaction). By incorporating both objective as well as subjective (self-reported) measures, evaluators can better understand the system's strengths and areas for improvement.

When designing an evaluation, the first step is to identify the stakeholders and their goals that need to be addressed. Based on the goals, the user tasks to be studied in the evaluation have to be defined, as well as the user aspects to be considered. The central element of an evaluation are the criteria to be focused on, which can be determined by the stakeholder goals. The chosen criteria restrict the range of possible evaluation methods (e. g. any user-centric criterion requires the involvement of actual users in the evaluation procedure). Finally, an appropriate measure has to be defined for any quantitative criterion.

2 Table of Contents

3 Overview of Talks

3.1 Conversational Search in 2019

Avishek Anand (TU Delft, NL, neil.hurley@ucd.ie)

In this talk we reflect on the results and insights from the last Dagstuhl Seminar in 2019 on conversational search (`https://www.dagstuhl.de/19461`). There are multiple definitions of conversational search systems or CSS and we looked at the Dagstuhl Typology. We also reflected on some of the challenges of the evaluation of CSS systems. Finally, we discussed about some potential open problems and challenges in the era of LLMs.

3.2 Preferences are Constructive: How to Build and Evaluate Better Conversational Interfaces that Really Give Guidance (with LLMs)?

Martijn C. Willemsen (TU Eindhoven, NL & JADS – 's-Hertogenbosch, NL, M.C.Willemsen@tue.nl)
Bart Knijnenburg (Clemson University, US, bartk@clemson.edu)

Recommender systems build user models to be able to predict users' preferences. However, preferences are volatile and often constructed while in the process of making decisions. In this talk we discuss ways in which recommender systems can go beyond just automatically providing recommendations to learn preferences better via active preference elicitation, interactive recommender systems and conversational interfaces such as critiquing-based recommender systems. We then discuss how such decision guidance should guide future developments of modern conversational agents including LLMs, and we discuss some of the pitfalls such as the persuasive nature and anthropomorphism that might users to over-trust such systems.

3.3 Conversational Recommenders: Reflecting on the Good, the Bad, and the Unknown

Maria Soledad Pera (TU Delft, NL, M.S.Pera@tudelft.nl)

In this talk, we take a somewhat provocative (or rather biased) approach to examining the differences (or lack thereof) between search and recommendation, and exploring what insights can be gained from research in these areas, particularly in terms of evaluation. We then provide a brief overview of studies on conversational recommenders published since the early 2000s, emphasizing evaluation perspectives. Finally, we discuss the challenges of evaluating conversational recommenders throughout different stages of the recommendation-generation process, including the choice of objectives to assess (simultaneously), the "right" metrics to use, data limitations, and how LLMs might increase the complexity of the evaluation process.

3.4 The Challenges and Opportunities in Evaluating Generative Information Retrieval

Mark Sanderson (RMIT University – Melbourne, AU, mark.sanderson@rmit.edu.au)

Evaluation has long been an important part of information retrieval research. Over decades of research, well established methodologies have been created and refined that for years have provided reliable relatively low cost benchmarks for assessing the effectiveness of retrieval systems. With the rise of generative AI and the explosion of interest in Retrieval Augmented Generation (RAG), evaluation is having to be rethought. In this talk, I will speculate on what might be solutions to evaluating RAG systems as well as highlighting some of the opportunities that are opening up. As important as it is to evaluate the new generative retrieval systems it is also important to recognize the traditional information retrieval has not yet gone away. However the way that these systems are being evaluated is undergoing a revolution. I will detail the transformation that is currently taking place in evaluation research. Here I will highlight some of the work that we've been doing at RMIT University as part of the exciting, though controversial, new research directions that generative AI is enabling.

Acknowledgements

We thank Schloss Dagstuhl for hosting us.

Participants

- Avishek Anand
TU Delft – Delft, NL
- Christine Bauer
Paris Lodron University – Salzburg, AT
- Timo Breuer
TH Köln, DE
- Li Chen
Hong Kong Baptist University, HK
- Guglielmo Faggioli
University of Padua, IT
- Nicola Ferro
University of Padua – Padova, IT
- Ophir Frieder
Georgetwon University – Washington, DC, US
- Norbert Fuhr
Universität Duisburg-Essen – Duisburg, DE
- Hideo Joho
University of Tsukuba – Ibaraki, JP
- Jussi Karlgren
Silo AI – Helsinki, FI
- Johannes Kiesel
Bauhaus-Universität Weimar, DE
- Bart Knijnenburg
Clemson University, SC, US
- Lien Michiels
imec-SMIT, Vrije Universiteit Brussel, BE & University of Antwerp, BE
- Andrea Papenmeier
University of Twente – Enschede, NL
- Maria Soledad Pera
TU Delft, NL
- Aldo Lipani
University College London, UK
- Mark Sanderson
RMIT University – Melbourne, AU
- Scott Sanner
University of Toronto, CA
- Benno Stein
Bauhaus-Universität Weimar, DE
- Johanne Trippas
RMIT University – Melbourne, AU
- Karin Verspoor
RMIT University – Melbourne, AU
- Martijn C. Willemsen
TU Eindhoven, NL & Jheronimus Academy of Data Science – 's-Hertogenbosch, NL