DAGSTUHL
REPORTS

Volume 15, Issue 6, June 2025

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
February, 2026

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.15.6.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 25231

Certifying Algorithms for Automated Reasoning

Nikolaj S. Bjørner[*1], Marijn J. H. Heule[*2], Daniela Kaufmann[*3], Jakob Nordström[*4], and Wietze Koops[†5]

1 **Microsoft - Redmond, US.** nbjorner@microsoft.com
2 **Carnegie Mellon University - Pittsburgh, US.** marijn@cmu.edu
3 **TU Wien, AT.** dk@danielakaufmann.at
4 **University of Copenhagen, DK & Lund University, SE.** jn@di.ku.dk
5 **Lund University, SE & University of Copenhagen, DK.** wietze.koops@cs.lth.se

Abstract

Modern automated reasoning has transformed large parts of industry and has also found numerous scientific applications. But many reasoning problems are computationally very challenging, or sometimes even undecidable. Because of this, the reasoning algorithms used are often very complex, and even the best current algorithms at times produce wrong results. As these tools are increasingly being used autonomously, sometimes even in life-critical applications, it is urgent to ensure that what they compute is valid. Software testing, while immensely useful, cannot guarantee correctness, and state-of-the-art algorithms are far beyond what techniques for producing formally verified software can handle.

The focus of this Dagstuhl Seminar was the approach of addressing such issues by designing certifying algorithms using so-called proof logging, meaning that algorithms output not only a result but also a machine-verifiable proof of correctness. This proof can then be fed to a dedicated proof checker for verification. Crucially, such proofs should require low overhead to generate and be easy to check, but still supply 100% correctness guarantees. Besides ensuring correctness of outputs for complex algorithms, proof logging can also provide new tools for algorithm development and analysis, software debugging, and even research into explainability in the context of AI.

Seminar June 1–6, 2025 – http://www.dagstuhl.de/25231

2012 ACM Subject Classification Mathematics of computing → Solvers; Theory of computation → Proof theory; Theory of computation → Automated reasoning; Software and its engineering → Software verification

Keywords and phrases ATP, Computer Algebra, DRAT, DRUP, MIP, Propagation Redundancy, QBF, SAT, SMT

Digital Object Identifier 10.4230/DagRep.15.6.1

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

Certifying Algorithms for Automated Reasoning, *Dagstuhl Reports*, Vol. 15, Issue 6, pp. 1–31
Editors: Nikolaj S. Bjørner, Marijn J. H. Heule, Daniela Kaufmann, Jakob Nordström, and Wietze Koops

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Nikolaj S. Bjørner (Microsoft - Redmond, US)
Marijn J. H. Heule (Carnegie Mellon University - Pittsburgh, US)
Daniela Kaufmann (TU Wien, AT)
Jakob Nordström (University of Copenhagen, DK & Lund University, SE)

Automated reasoning has been widely adopted over the last decades for developing formally verified software and also in the context of combinatorial optimization. The foundation is built on automated deduction algorithms that are used for determining the satisfiability of propositional logic, first-order logic, and arithmetic formulas.

Algorithms for determining the validity of Boolean satisfiability (SAT), mixed integer programming (MIP), satisfiability modulo theories (SMT), and first-order automated theorem proving (ATP) formulas are integral components in verification tools. The question of how to certify correctness of the conclusions reached by such reasoning algorithms has received long-running attention, since in industrial formal verification they form the trusted base for correctness of safety-critical applications such as control systems for trains and airplanes.

Another application of automated reasoning is in combinatorial optimization, which studies problems where solutions are constructed by combining objects, but where the supply of objects is limited and there are constraints how they can be utilized together. Combinatorial optimization problems are encountered in a multitude of practical scenarios including logistics, scheduling, and disaster management. Here reasoning algorithms are employed not only to establish when a feasible solution is separated from non-feasible bounds, but they are also deeply integrated within combinatorial optimization solvers that can provide guaranteed optimal answers. The computational complexity of solving combinatorial optimization problems, which are typically harder than *NP*-complete, as well as the complexity of implementing sophisticated combinatorial solvers in practice, provide major challenges.

Algorithms used in symbolic solvers are often stunningly powerful in practice, and are today used routinely to solve large-scale real-world problems in a wide range of application areas. But the "dirty little secret" is that the solvers are sometimes wrong. It is well documented that even the best constraint programming (CP) and mixed integer programming (MIP) solvers sometimes return "solutions" that do not satisfy the constraints, or erroneously claim optimality, and that verification tools can erroneously claim a set of constraints is infeasible when, on the contrary, it has a solution. Also, in more complex scenarios, where solvers are used to solve subproblems, even seemingly innocuous off-by-one mistakes can snowball into huge overall errors.

Dealing with errors in software is, of course, not a new problem. The traditional method to discover and eliminate bugs is *software testing*. However, while substantial progress has been made recently on powerful so-called *fuzzing-based tools* applied to symbolic solvers, they cannot offer any guarantees that results produced by a solver are correct. It is an inherent limitation that testing can only reveal the presence of bugs, but never prove their absence.

Another very appealing approach is *formal verification*. This means that one writes down a formal, mathematical, specification of how the solver should work, and then provides a proof that the solver adheres to this specification. The main obstacle for this method is that the advanced techniques used in state-of-the-art solvers go far beyond what formal verification can currently handle. And even a fully verified solver cannot deal with the problem of incorrect results arising from hardware failures, which are unavoidable in large-scale computations.

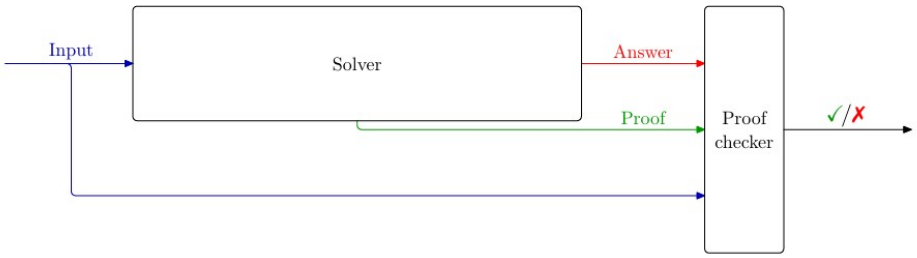

Figure 1 Schematic workflow for solver with proof logging.

Thus, the state of the art when it comes to verifiably correct automated reasoning is that this is a well-recognized problem that has remained without convincing solutions.

The Main Focus of This Seminar

This seminar focused on what currently appears to be the most promising way to eliminate errors in automated reasoning algorithms, namely *proof logging*. This means turning solvers into *certifying algorithms* in the sense of [1, 12] by having them output not only an answer but also a simple, machine-verifiable *proof* that this answer is correct. With such a solver, the workflow becomes as follows (see also Figure 1):

1. Run the solver on a problem to obtain an answer together with a proof.
2. Feed the problem, the answer, and the proof to a special computer program, called a *proof checker*.
3. Accept the answer if the proof checker says that the proof is valid.

For this to be feasible, the proof format needs to be sufficiently powerful, so that the solver can generate concise proofs even for sophisticated reasoning without incurring any serious overhead in running time. But the proof format should also be very simple, so that checking correctness becomes almost trivial – the point is that the proof checker, in contrast to the solver, should be so easy to code up that we can be confident that it is correct. Clearly, there is a conflict between expressivity and simplicity here. Perhaps asking for both at the same time is a little bit too good to be true? This tension between succinct proofs and easy verification goes to the heart of proof logging, and discussions of different ways of managing this trade-off was one of the key topics of the seminar.

One example of an approach that has so far been found to be unsatisfactory are methods in constraint programming using *explanations* [13, 15, 5], which essentially boils down to writing out reasons for solver conclusions trusting that these reasons are correct. The problem is that this means that proofs are so expressive that they cannot be efficiently verified by simple proof checking algorithms.

A much more successful approach is the *DRAT* proof system [9, 8, 16], which has become standard in SAT solving. This proof system is simple enough that the proof checker can even be implemented as a formally verified piece of software [4, 11], meaning that the full power of formal methods can be harnessed to guarantee correctness of the result produced by the solving algorithm. The crucial change of perspective making this possible is that the guarantee is not that the *algorithm* is correct, but that the *answer* found by the algorithm is.

And, in some sense, this even goes further than formally verified software in two important ways. Firstly, formally verified proof checking makes it possible to detect errors even if they are not due to faults in the solver but are caused by a buggy compiler, faulty hardware, or even cosmic rays during solver execution. Secondly, formally verified proof checking can allow us to fully trust the results even from buggy solvers. If the proof generated by a concrete algorithm execution checks out, then we can be fully confident that this particular computation reached the correct result, regardless of what bugs might be triggered for other inputs. It seems fair to argue that for reasons like this, SAT proof logging is perhaps the most successful showcase of certifying algorithms for computationally challenging problems to date, and for this reason it was natural to survey this area and discuss how similar advances could be made in other areas of automated reasoning.

However, when one tries to extend proof logging to stronger optimization paradigms such as *maximum satisfiability (MaxSAT)*, *pseudo-Boolean optimization*, *CP*, and *MIP*, or other areas of automated reasoning such as *automated planning* and *SMT solving*, the conflict between expressivity and simplicity reasserts itself. The clean and efficient reasoning in terms of disjunctive clauses used in *DRAT* seems poorly suited to capture reasoning about more complex objects like constraint programming propagators. Some of the other reasoning performed in more advanced solvers also seems hard to express in terms of the disjunctive clauses used in *DRAT*, and this limitation also makes it hard argue about, e.g., values of objective functions in optimization problems. At a high-level, the reason that this is a highly nontrivial challenge is that the stronger the solving techniques used, the harder it becomes to design simple proof logging methods that can efficiently certify the correctness of these techniques. At the same time, the fact that CP, MIP, and SMT solvers are so fiendishly complex only makes the need for proof logging methods in these settings even more urgent.

In SMT solving, the most popular approach to date seems to be to design very expressive proof systems that can capture all the different theories considered and their combinations. One downside with this that the proof systems become extremely complex, meaning that not only does the proof checking algorithm have to be very involved, but it is even a highly nontrivial task to even decide whether the proof system itself is consistent. Another direction, which has recently been pursued in the context of CP proof logging, is to compile all information about the input problem down to a simpler format in a trusted (or formally verified) way, and then mirror the reasoning performed in the solver by a proof in this simpler format. During the seminar, different subcommunities in automated reasoning were able to exchange experiences and best practices for these and other proof logging approaches for automated reasoning paradigms beyond SAT.

Further Topics Discussed During the Seminar

While the initial motivation for proof logging techniques is that they provide a way of ensuring the validity of outputs from complex algorithms, discussions at the seminar ranged over a number of topics that went beyond just providing formal **certificates of correctness** for answers produced by automated reasoning algorithms. Several participants of the seminar highlighted that proof logging can also be used as a tool for **debugging** during software development. Bugs that only very rarely affect the final result can be next to impossible to discover, but with proof logging switched on, it is easy to detect that the algorithm is performing unsound reasoning even when the output happens to be correct (as shown in, e.g., [6, 7, 10, 3]). It is also worth noting that it simplifies test case generation during debugging. There is no need to know what the correct output is, and instead testing can be done by checking the proof log.

Designing proof logging for a concrete algorithm typically involves describing in a formal proof system how the algorithm works, so that different reasoning steps can be written down as rule applications in the proof system. This type of analysis can also be quite helpful for **algorithm design** in that it can identify limitations in the current implementation of an algorithm and uncover potential for further improvements (if the proof system suggest that more powerful reasoning steps could be applied than what the algorithm actually does). Furthermore, since proof logging allows us to "peek inside" the algorithm, as it were, to get detailed information about what reasoning steps were performed, this provides a new tool for in-depth, scientifically rigorous, **performance analysis**.

Going further, it can be noted that proof logging by its very nature enables **auditability**, since once an algorithm execution has finished we can save the problem, answer, and proof for posterity so that it can be verified at any time by a third party, even if this third party has no access to the original algorithm used to solve the problem Also, the fact that a proof for an optimization problem shows in a formal, mathematical, sense why a solution is correct, and/or why no better solution is possible, means that proof logging can serve as a stepping stone towards **explainability**, which is a topic of growing importance in the context of artificial intelligence (AI). These and other aspects were discussed in presentation by participants and also during a dedicated panel discussion towards the end of the seminar week.

The topics outlined above are well-represented in the research pursued by the researchers who participated in the seminar. In the SAT community, Armin Biere, Katalin Fazekas, Mathias Fleury, Marijn Heule, Adrián Rebola-Pardo and others have developed proof systems based on *DRAT* and extensions. Efficient and formally verified proof checkers for such proof systems have been built by Magnus Myreen and Yong Kiam Tan. Proof logging techniques for combinatorial optimization paradigms beyond SAT have been successfully pursued by Jeremias Berg, Bart Bogaerts, Wietze Koops, Ciaran McCreesh, Matthew McIlree, Jakob Nordström, Andy Oertel, and their collaborators. For mixed integer programming, Ambros Gleixner has done important exploratory work on proof logging for perhaps the most well-known open-source MIP solver *SCIP* together with collaborators. Certification of SMT solvers has received long-running attention, including work on *Z3* by Nikolaj Bjørner and *cvc5* by Haniel Barbosa, Bruno Dutertre, Hanna Lachnitt, and Andrew Reynolds together with collaborators. Many of these researchers gave presentations of their work where they also identified important future challenges.

In the context of automated theorem proving (ATP) for first-order logic, there are natural connections between formats for certifying deductions of ATP systems on the one hand and proof logging and model verification on the other hand, but also formidable technical (and organizational) obstacles to wider adoption of such techniques. During the seminar, Geoff Sutcliffe and Michael Rawson presented new results on ATP proof logging and checking. The participants also provided coverage of research on algebraic algorithms (Daniela Kaufmann), quantified Boolean formula solving (Martina Seidl), automated planning (Malte Helmert and Tanja Schindler), hardware verification (Dirk Beyer and Randal Bryant), hybrid systems (Erika Ábrahám), and several other topics related to automated reasoning.

Outcomes

The seminar aimed to advance the state of the art in the integration of proof logging with symbolic solvers, and to establish deeper contacts between different research communities working on certifying algorithms where interaction has previously been quite limited or non-

existent. The intention was to achieve this broad goal by assembling stakeholders in Boolean satisfiability (SAT) solving, constraint programming (CP), Mixed integer programming (MIP), satisfiability modulo theories (SMT) solving, automated theorem proving (ATP), and other closely related communities, including leading researchers in the areas of solver development, deployment of solver tools in applications, and design of proof logging techniques. Concretely, the seminar aimed to:

1. Connect automated-reasoning experts from the different domains around proof logging techniques.
2. Infuse the communities with new insights into the practical integration of proof logging and methods to develop formally verified proof checkers.
3. Facilitate technology transfer between different research areas in automated reasoning, in particular, concerning techniques for certifying correctness.

Going by the evaluations, the seminar was very successful in reaching these goals. It is our hope that this seminar will turn out to be only the first in a series of seminars dedicated to the important topic of certifying algorithms for automated reasoning. In the longer perspective, our vision is that such a series of seminars would contribute to a fundamental shift in how the computer science community thinks about algorithms, so that in the future algorithms will be expected to not just produce output but to prove that this output is in fact correct.

Seminar Structure

The scientific program of the seminar consisted of 30 presentations. Among these there were eleven 50-minute surveys of different core topics of the seminar. These talks occupied most of the morning schedule Monday-Wednesday, and were intended to make sure that the diverse audience would have a bit of a common background for the more technical talks reporting on recent progress and/or ongoing research. The list of survey talks and speakers were as follows:

- Certified SAT solving (Katalin Fazekas)
- Certified subgraph solving (Ciaran McCreesh)
- Certified constraint programming (Matthew McIlree)
- Proof logging for algebraic algorithms (Daniela Kaufmann)
- Proof logging for MIP (Ambros Gleixner)
- Certified automated planning (Malte Helmert)
- Certified SMT solving (Haniel Barbosa)
- Certified model counting and knowledge compilation (Randal Bryant)
- Certified QBF solving (Martina Seidl)
- Certified first-order theorem proving (Michael Rawson)
- Formally verified proof checking (Magnus O. Myreen & Yong Kiam Tan)

The rest of the talks were 25-minute presentations on recent research of the participants. The time after lunch each day was left for self-organized collaborations and discussions, and there was no schedule on Wednesday afternoon.

Based on polling of participants during the seminar, it was decided to have a panel discussion on Thursday afternoon. The poll also asked whether an open-problem session should be organized, but the support for this idea was weaker, and several participants emphasized that the seminar program should not be too dense and that the evenings should be left free of any program. Therefore, the organizers decided not to have an open-problem session.

References

1 Eyad Alkassar, Sascha Böhme, Kurt Mehlhorn, Christine Rizkallah, and Pascal Schweitzer. An introduction to certifying algorithms. *it - Information Technology Methoden und innovative Anwendungen der Informatik und Informationstechnik*, 53(6):287–293, December 2011.

2 Haniel Barbosa, Clark W. Barrett, Byron Cook, Bruno Dutertre, Gereon Kremer, Hanna Lachnitt, Aina Niemetz, Andres Nötzli, Alex Ozdemir, Mathias Preiner, Andrew Reynolds, Cesare Tinelli, and Yoni Zohar. Generating and exploiting automated reasoning proof certificates. *Commun. ACM*, 66(10):86–95, 2023.

3 Jeremias Berg, Bart Bogaerts, Jakob Nordström, Andy Oertel, and Dieter Vandesande. Certified core-guided MaxSAT solving. In *Proceedings of the 29th International Conference on Automated Deduction (CADE-29)*, volume 14132 of *Lecture Notes in Computer Science*, pages 1–22. Springer, July 2023.

4 Luís Cruz-Filipe, Marijn J. H. Heule, Warren A. Hunt Jr., Matt Kaufmann, and Peter Schneider-Kamp. Efficient certified RAT verification. In *Proceedings of the 26th International Conference on Automated Deduction (CADE-26)*, volume 10395 of *Lecture Notes in Computer Science*, pages 220–236. Springer, August 2017.

5 Nicholas Downing, Thibaut Feydy, and Peter J. Stuckey. Explaining alldifferent. In *Proceedings of the 35th Australasian Computer Science Conference (ACSC '12)*, pages 115–124, January 2012.

6 Leon Eifler and Ambros Gleixner. A computational status update for exact rational mixed integer programming. In *Proceedings of the 22nd International Conference on Integer Programming and Combinatorial Optimization (IPCO '21)*, volume 12707 of *Lecture Notes in Computer Science*, pages 163–177. Springer, May 2021.

7 Stephan Gocht, Ross McBride, Ciaran McCreesh, Jakob Nordström, Patrick Prosser, and James Trimble. Certifying solvers for clique and maximum common (connected) subgraph problems. In *Proceedings of the 26th International Conference on Principles and Practice of Constraint Programming (CP '20)*, volume 12333 of *Lecture Notes in Computer Science*, pages 338–357. Springer, September 2020.

8 Marijn J. H. Heule, Warren A. Hunt Jr., and Nathan Wetzler. Trimming while checking clausal proofs. In *Proceedings of the 13th International Conference on Formal Methods in Computer-Aided Design (FMCAD '13)*, pages 181–188, October 2013.

9 Marijn J. H. Heule, Warren A. Hunt Jr., and Nathan Wetzler. Verifying refutations with extended resolution. In *Proceedings of the 24th International Conference on Automated Deduction (CADE-24)*, volume 7898 of *Lecture Notes in Computer Science*, pages 345–359. Springer, June 2013.

10 Sonja Kraiczy and Ciaran McCreesh. Solving graph homomorphism and subgraph isomorphism problems faster through clique neighbourhood constraints. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI '21)*, pages 1396–1402, August 2021.

11 Peter Lammich. Efficient verified (UN)SAT certificate checking. *Journal of Automated Reasoning*, 64(3):513–532, March 2020. Extended version of paper in *CADE* 2017.

12 Ross M. McConnell, Kurt Mehlhorn, Stefan Näher, and Pascal Schweitzer. Certifying algorithms. *Computer Science Review*, 5(2):119–161, May 2011.

13 Olga Ohrimenko, Peter J. Stuckey, and Michael Codish. Propagation via lazy clause generation. *Constraints*, 14(3):357–391, January 2009.

14 Alexander Steen, Geoff Sutcliffe, Pascal Fontaine, and Jack McKeown. Representation, verification, and visualization of Tarskian interpretations for typed first-order logic. In *LPAR 2023: Proceedings of 24th International Conference on Logic for Programming,*

Artificial Intelligence and Reasoning, volume 94 of *EPiC Series in Computing*, pages 369–385. EasyChair, June 2023.

15 Michael Veksler and Ofer Strichman. A proof-producing CSP solver. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI '10)*, pages 204–209, July 2010.

16 Nathan Wetzler, Marijn J. H. Heule, and Warren A. Hunt Jr. DRAT-trim: Efficient checking and trimming using expressive clausal proofs. In *Proceedings of the 17th International Conference on Theory and Applications of Satisfiability Testing (SAT '14)*, volume 8561 of *Lecture Notes in Computer Science*, pages 422–429. Springer, July 2014.

2 Table of Contents

3 Overview of Talks

3.1 The certification problem for real algebra

Erika Ábrahám (RWTH Aachen University, DE)

Joint work of Erika Ábrahám, Jasper Nalbach, Valentin Promies

SMT solvers' traditional functionality is to check the satisfiability of quantifier-free formulas of first-order logic over different theories.

With their increasing efficiency and usage, this original functionality is being extended in different directions. One of them is the ability to provide some kind of assurance for the correctness of the computations, most prominently in the form of certificates.

Whereas some SMT solvers can already provide certificates for a wide range of theories, the theory of (quantifier-free non-linear) real algebra poses a hard challenge, and a solution seems to be yet completely out of reach.

In this talk, we discussed why this problem is especially hard, and which directions could be considered to make some progress.

3.2 Symbolic Conflict Analysis in Pseudo-Boolean Solving

Albert Oliveras (UPC Barcelona Tech, ES)

Joint work of Robert Nieuwenhuis, Albert Oliveras, Enric Rodríguez-Carbonell, Rui Zhao
Main reference Robert Nieuwenhuis, Albert Oliveras, Enric Rodríguez-Carbonell, Rui Zhao: "Symbolic Conflict Analysis in Pseudo-Boolean Optimization", in Proc. of the 28th International Conference on Theory and Applications of Satisfiability Testing, SAT 2025, August 12-15, 2025, Glasgow, Scotland, LIPIcs, Vol. 341, pp. 23:1–23:18, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org//10.4230/LIPICS.SAT.2025.23

In the last two decades, a lot of effort has been devoted to the development of satisfiability-checking tools for a variety of SAT-related problems. However, most of these tools lack optimization capabilities. That is, instead of finding any solution, one is sometimes interested in a solution that is best according to some criterion.

Pseudo-Boolean solvers can be used to deal with optimization by successively solving a series of problems that contain an additional pseudo-Boolean constraint expressing that a better solution is required. A key point for the success of this simple approach is that lemmas that are learned for one problem can be reused for subsequent ones.

In this talk we go one step further and show how, by using a simple symbolic conflict analysis procedure, not only can lemmas be reused between problems but also strengthened, thus further pruning the search space traversal. In addition, we show how this technique automatically allows one to infer upper bounds in maximization problems, thus giving an estimation of how far the solver is from finding an optimal solution. Experimental results with our PB solver reveal that (i) this technique is indeed effective in practice, providing important speedups in problems where several solutions are found and (ii) on problems with very few solutions, where the impact of our technique is limited, its overhead is negligible.

3.3 Faster Certified Symmetry Breaking in SAT

Markus Anders (RPTU Kaiserslautern-Landau, DE)

Joint work of Markus Anders, Bart Bogaerts, Benjamin Bogø, Arthur Gontier, Wietze Koops, Ciaran McCreesh, Magnus Myreen, Jakob Nordström, Andy Oertel, Adrián Rebola-Pardo, Yong Kiam Tan

Symmetry breaking is a standard technique in many areas of automated reasoning. Recently, the possibility for proof logging symmetry breaking techniques in SAT solvers has become available by means of the dominance rule and VeriPB proof system [2]. It turns out however, that the proposed logging and checking techniques pose a severe bottleneck for efficient, modern symmetry handling algorithms [1]. In this talk, I gave a brief overview of symmetry handling algorithms and related proof logging techniques. Then, I discussed recent developments to improve logging and checking performance through the introduction of auxiliary variables. Lastly, I mentioned some of the remaining challenges.

References

1 Markus Anders, Sophie Brenner, and Gaurav Rattan. *Satsuma: Structure-Based Symmetry Breaking in SAT*. In Proc. of the 27th International Conference on Theory and Applications of Satisfiability Testing, SAT 2024, August 21-24, 2024, Pune, India, LIPIcs, Vol. 305, pp. 4:1–4:23, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.
https://doi.org//10.4230/LIPICS.SAT.2024.4

2 Bart Bogaerts, Stephan Gocht, Ciaran McCreesh, Jakob Nordström. *Certified Dominance and Symmetry Breaking for Combinatorial Optimisation*. J. Artif. Intell. Res., Vol. 77, 2023.
https://doi.org//10.1613/JAIR.1.14296

3.4 First Results on How to Certify Subsumptions Computed by the EL Reasoner Elk Using the Logical Framework with Side Conditions

Franz Baader (TU Dresden, DE)

Joint work of Franz Baader, Patrick Koopmann, Cesare Tinelli
Main reference Franz Baader, Patrick Koopmann, Cesare Tinelli: "First Results on How to Certify Subsumptions Computed by the EL Reasoner ELK Using the Logical Framework with Side Conditions", in Proc. of the 33rd International Workshop on Description Logics (DL 2020) co-located with the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020), Online Event [Rhodes, Greece], September 12th to 14th, 2020, CEUR Workshop Proceedings, Vol. 2663, CEUR-WS.org, 2020.
URL https://ceur-ws.org/Vol-2663/paper-5.pdf

The generation of proof certificates and the use of proof checkers is nowadays standard in first-order automated theorem proving and related areas. They have, to the best of our knowledge, not yet been employed in Description Logics, where the focus was on detecting and repairing errors in the ontology, rather than on catching erroneous consequences created by an incorrect reasoner. This paper reports on first steps towards remedying this deficit for subsumptions computed by the DL reasoner Elk. We use an existing tool for generating proofs of consequences from Elk, and transform these proofs into a format that is accepted as certificates by our proof checker. The checker is obtained as an instance of a generic certification tool based on the Logical Frame-work with Side Conditions (LFSC), by formalizing the inference rules of Elk in LFSC. We report on the results of applying this approach to the classification of a large number of real-world OWL 2 EL ontologies.

3.5 SMT proof production, checking and reconstruction

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

Main reference Haniel Barbosa, Clark W. Barrett, Byron Cook, Bruno Dutertre, Gereon Kremer, Hanna Lachnitt, Aina Niemetz, Andres Nötzli, Alex Ozdemir, Mathias Preiner, Andrew Reynolds, Cesare Tinelli, Yoni Zohar: "Generating and Exploiting Automated Reasoning Proof Certificates", Commun. ACM, Vol. 66(10), pp. 86–95, 2023.
URL https://doi.org//10.1145/3587692

Main reference Haniel Barbosa, Andrew Reynolds, Gereon Kremer, Hanna Lachnitt, Aina Niemetz, Andres Nötzli, Alex Ozdemir, Mathias Preiner, Arjun Viswanathan, Scott Viteri, Yoni Zohar, Cesare Tinelli, Clark Barrett: "Flexible proof production in an industrial-strength SMT solver," in Proc. of the 11th International Joint Conference on Automated Reasoning, IJCAR 2022, August 8-10, 2022, Haifa, Israel, LNCS, Vol. 13385, pp. 15–35, Springer, 2022.
URL https://doi.org//10.1007/978-3-031-10769-6_3

Main reference Abdalrhman Mohamed, Tomaz Mascarenhas, Harun Khan, Haniel Barbosa, Andrew Reynolds, Yicheng Qian, Cesare Tinelli, Clark Barrett: "Lean-SMT: An SMT tactic for discharging proof goals in Lean". CoRR, Vol. abs/2505.15796, 2025.
URL https://doi.org//10.48550/ARXIV.2505.15796

SMT solvers can be hard to trust, since it generally means assuming their large and complex codebases do not contain bugs leading to wrong results. Machine-checkable certificates, via proofs of the logical reasoning the solver has performed, address this issue by decoupling confidence in the results from the solver's implementation. In this talk we will describe the extensive proof infrastructure of the state-of-the-art SMT solver *cvc5*, which has enabled the production of proofs in a number of complex domains. We will also show how these proofs are checked or reconstructed in different formats by different systems, from ad-hoc high-performance proof checkers to proof assistants such as Lean.

3.6 Certifying Software Verification

Dirk Beyer (LMU München, DE)

Joint work of Paulína Ayaziová, Dirk Beyer, Marian Lingsch-Rosenfeld, Martin Spiessl, Jan Strejček
Main reference Paulína Ayaziová, Dirk Beyer, Marian Lingsch-Rosenfeld, Martin Spiessl, Jan Strejček: "Software Verification Witnesses 2.0", in Proc. of the Model Checking Software - 30th International Symposium, SPIN 2024, Luxembourg City, Luxembourg, April 8-9, 2024, Proceedings, Lecture Notes in Computer Science, Vol. 14624, pp. 184–203, Springer, 2024.
URL https://doi.org//10.1007/978-3-031-66149-5_11

Over the last years, certifying software verification has become an established practice in the area of automatic software verification: An independent validator re-establishes verification results of a software verifier using verification certificates (also called witnesses), which are stored in a standardized exchange format. In addition to validation, such exchangeable information about proofs and alarms found by a verifier can be shared across verification tools, and users can apply independent third-party tools to visualize and explore certificates to help them comprehend the causes of bugs or the reasons why a given program is correct. To achieve the goal of making verification results more accessible to engineers, it is necessary to consider certificates as first-class exchangeable objects, stored independently from the source code and checked independently from the verifier that produced them, respecting the important principle of separation of concerns. We present the conceptual principles of software-verification certificates and illustrate the contents of such certificates.
Material:

- Software Verification Witnesses 2.0 https://doi.org/10.1007/978-3-031-66149-5_11
- Verification Witnesses https://doi.org/10.1145/3477579

References

1 Paulína Ayaziová, Dirk Beyer, Marian Lingsch-Rosenfeld, Martin Spiessl, Jan Strejček. *Software verification witnesses 2.0.* In Proc. of the 30th International Symposium on Model Checking Software, SPIN 2024, April 8-9, 2024, Luxembourg City, Luxembourg, LNCS, Vol. 14624, pp. 184–203, Springer, 2024. https://doi.org//10.1007/978-3-031-66149-5_11

2 Dirk Beyer, Matthias Dangl, Daniel Dietsch, Matthias Heizmann, Thomas Lemberger, Michael Tautschnig: *Verification witnesses.* ACM Trans. Softw. Eng. Methodol, Vol. 31(4), pp. 57:1–57:69, 2022. https://doi.org//10.1145/3477579

3.7 Certifying Hardware Model Checking

Armin Biere (Universität Freiburg, DE)

Joint work of Nils Froleyks, Emily Yu, Mathias Preiner, Armin Biere, Keijo Heljanko
Main reference Nils Froleyks, Emily Yu, Mathias Preiner, Armin Biere, Keijo Heljanko: "Introducing Certificates to the Hardware Model Checking Competition", in Proc. of the Computer Aided Verification: 37th International Conference, CAV 2025, Zagreb, Croatia, July 23-25, 2025, Proceedings, Part I, p. 281–295, Springer-Verlag, 2025.
URL https://doi.org//10.1007/978-3-031-98668-0_14

Design faults in hardware design are costly. Thus hardware model checking has routinely been applied during the chip design process for decades. However, both academic and industrial model checkers are complex software tools and arguably hard to get correct. To increase trust in model checkers we therefore propose a model checking certification flow. The model checker produces a witness circuit which simulates the original model and for safety properties has an inductive property implying the original property. Checking simulation and inductiveness can be done by SAT solving. We have applied this idea to different model checking techniques, particularly preprocessing techniques. The single safety property track of the hardware model checking competition in 2024 required all participants to produce such certificates. The competition showed that certification is possible and cheap, i.e., both with respect to certificate production and checking. Furthemore the winner of the competition surpases the previous state-of-the-art, while producing machine checked witnesses. This is joint work with Emily Yu, Nils Froleyks, Mathias Preiner and Keijo Heljanko.

3.8 Certifying Pareto Optimality in Multi-Objective Maximum Satisfiability

Bart Bogaerts (KU Leuven, BE)

Joint work of Christoph Jabs, Bart Bogaerts, Jeremias Berg, Matti Järvisalo
Main reference Christoph Jabs, Jeremias Berg, Bart Bogaerts, Matti Järvisalo: "Certifying Pareto-Optimality in Multi Objective Maximum Satisfiability", in Proc. of the Tools and Algorithms for the Construction and Analysis of Systems - 31st International Conference, TACAS 2025, Held as Part of the International Joint Conferences on Theory and Practice of Software, ETAPS 2025, Hamilton, ON,

Canada, May 3-8, 2025, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 15697, pp. 108–129, Springer, 2025.
URL https://doi.org//10.1007/978-3-031-90653-4_6

Due to the wide employment of automated reasoning in the analysis and construction of correct systems, the results reported by automated reasoning engines must be trustworthy. For Boolean satisfiability (SAT) solvers – and more recently SAT-based maximum satisfiability (MaxSAT) solvers – trustworthiness is obtained by integrating proof logging into solvers, making solvers capable of emitting machine-verifiable proofs to certify correctness of the reasoning steps performed. In this work, we enable for the first time proof logging based on the VeriPB proof format for multi-objective MaxSAT (MO-MaxSAT) optimization techniques. Although VeriPB does not offer direct support for multiobjective problems, we detail how preorders in VeriPB can be used to provide certificates for MO-MaxSAT algorithms computing a representative solution for each element in the non-dominated set of the search space under Pareto-optimality, without extending the VeriPB format or the proof checker. By implementing VeriPB proof logging into a state-of-the-art multi-objective MaxSAT solver, we show empirically that proof logging can be made scalable for MO-MaxSAT with reasonable overhead.

3.9 Checkable Proofs for Model Counting and Knowledge Compilation

Randal E. Bryant (Carnegie Mellon University - Pittsburgh, US)

Joint work of Wojciech Nawrocki, Jeremy Avigad, Randal E. Bryant, Yong Kiam Tan, Marijn J. H. Heule
Main reference Randal E. Bryant, Yong Kiam Tan, Marijn J. H. Heule: "Certifying Projected Knowledge Compilation", in Proc. of the 28th International Conference on Theory and Applications of Satisfiability Testing (SAT 2025), Leibniz International Proceedings in Informatics (LIPIcs), Vol. 341, pp. 8:1–8:22, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org//10.4230/LIPIcs.SAT.2025.8

Knowledge compilers convert Boolean formulas, given in conjunctive normal form (CNF), into representations that enable efficient evaluation of unweighted and weighted model counts, as well as a variety of other useful properties. Certifying the correctness of a knowledge compiler's output, requires proving that 1) the generated formula is logically equivalent to the input formula, and 2) the generated formula satisfies the structural properties that enable efficient model counting.

Our Certified Partitioned-Operation Graph (CPOG) proof framework provides a way to encode the output of a knowledge compiler as well as a set of steps providing a checkable proof of correctness. Most recently, we have extended this framework to Skolem CPOG (SCPOG) supporting projected knowledge compilation, where a subset of the variables is abstracted away via existential quantification. Doing so requires a method to encode Skolem assignments, describing instantiations of the quantified variables.

We have developed formally verified checkers for both CPOG and SCPOG, one in Lean4 and the other in CakeML/HOL. In doing so, we formally verified the soundness of the frameworks.

3.10 Certified SAT solving

Katalin Fazekas (TU Wien, AT)

Joint work of Katalin Fazekas, Florian Pollitt, Mathias Fleury, Armin Biere
Main reference Katalin Fazekas, Florian Pollitt, Mathias Fleury, Armin Biere: "Certifying Incremental SAT Solving", in Proc. of the LPAR 2024: Proceedings of 25th Conference on Logic for Programming, Artificial Intelligence and Reasoning, Port Louis, Mauritius, May 26-31, 2024, EPiC Series in Computing, Vol. 100, pp. 321–340, EasyChair, 2024.
URL https://doi.org//10.29007/PDCC

This invited survey talk explores certified SAT solving, which is crucial for establishing trust in the reasoning steps and results of Boolean Satisfiability (SAT) solvers. We will cover the related fundamental concepts of SAT solving and discuss how proof-producing solvers and external checkers enable certification. A key focus will be on certifying incremental SAT solving, an essential technique that allows solvers to efficiently tackle sequences of related problems while maintaining correctness guarantees.

3.11 Consuming CaDiCaL Proofs

Mathias Fleury (Universität Freiburg, DE)

Joint work of Armin Biere, Tobias Faller, Katalin Fazekas, Mathias Fleury, Nils Froleyks, Florian Pollitt
Main reference Armin Biere, Tobias Faller, Katalin Fazekas, Mathias Fleury, Nils Froleyks, Florian Pollitt: "CaDiCaL 2.0", in Proc. of the Computer Aided Verification - 36th International Conference, CAV 2024, Montreal, QC, Canada, July 24-27, 2024, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 14681, pp. 133–152, Springer, 2024.
URL https://doi.org//10.1007/978-3-031-65627-9_7

CaDiCaL offers a proof tracer interface that makes it possible to get information on the derivation worked. This interface hides only some of the internal information. In this talk, I describe the notification model and what information is produced.

3.12 Proof logging and proof production for Mixed-Integer Programming

Ambros Gleixner (HTW - Berlin, DE)

Joint work of Ambros Gleixner, Leon Eifler, Alexander Hoen

Standard solvers for mixed-integer linear programming define feasibility and optimality of solutions within numerical tolerances and the correctness of their results, even within these tolerances, is subject to roundoff errors stemming from the unsafe use of floating-point arithmetic. By contrast, starting with version 10, the open-source MIP solver SCIP ships a numerically exact solving mode without tolerances and can produce an independently verifiable proof log for most of the exact solving techniques. Besides giving an overview on these recent advances and remaining limitations in software for verified MIP solving, we try to gauge to what extent floating-point MIP solvers can be used directly to produce verifiably correct proof logs. Our computational study with a pure LP-based branch-and-bound version

of SCIP confirms the expectation that in the overwhelming majority of cases, all critical decisions during the solving process are correct. When errors do occur on numerically challenging instances, they typically affect only a small, typically single-digit, amount of leaf nodes that would require further processing.

References

1 Leon Eifler and Ambros Gleixner. *Safe and verified Gomory mixed-integer cuts in a rational mixed-integer program framework.* SIAM Journal on Optimization, Vol. 34(1), pp. 742–763, 2024. https://doi.org//10.1137/23M156046X

2 Alexander Hoen and Ambros Gleixner. *Analyzing the numerical correctness of branch-and-bound decisions for mixed-integer programming.* In Proc. of the 22nd International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research, CPAIOR 2025, November 10–13, 2025, Melbourne, Victoria, Australia, LNCS, Vol. 15763, pp. 35–50, Springer, 2025. https://doi.org//10.1007/978-3-031-95976-9_3

3.13 Speculative SAT modulo SAT

Arie Gurfinkel (University of Waterloo, CA)

Joint work of Arie Gurfinkel, Hari Govind VK, Isabel Garcia-Contreras, Sharon Shoham
Main reference Hari Govind V. K., Isabel Garcia-Contreras, Sharon Shoham, Arie Gurfinkel: "Speculative SAT Modulo SAT", in Proc. of the Tools and Algorithms for the Construction and Analysis of Systems - 30th International Conference, TACAS 2024, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2024, Luxembourg City, Luxembourg, April 6-11, 2024, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 14570, pp. 43–60, Springer, 2024.
URL https://doi.org//10.1007/978-3-031-57246-3_4

State-of-the-art model-checking algorithms like IC3/PDR are based on unidirectional modular SAT solving for finding and/or blocking counterexamples. Modular SAT solvers divide a SAT-query into multiple sub-queries, each solved by a separate SAT solver (called a module), and propagate information (lemmas, proof obligations, blocked clauses, etc.) between modules. While modular solving is key to IC3/PDR, it is obviously not as effective as monolithic solving, especially when individual sub-queries are harder to solve than the combined query. This is partially addressed in SAT modulo SAT (SMS) by propagating unit literals back and forth between the modules and using information from one module to simplify the sub-query in another module as soon as possible (i.e., before the satisfiability of any sub-query is established). However, bi-directionality of SMS is limited because of the strict order between decisions and propagation – only one module is allowed to make decisions, until its sub-query is SAT. In this talk, I will describe our generalization of SMS, called SPEC SMS, that speculates decisions between modules. This makes it bi-directional – decisions are made in multiple modules, and learned clauses are exchanged in both directions. We further extend DRUP proofs and interpolation, these are useful in model checking, to SPEC SMS. We have implemented SPEC SMS in Z3 and show that it performs exponentially better on a series of benchmarks that are provably hard for SMS.

3.14 Certified Automated Planning

Malte Helmert (Universität Basel, CH)

In my talk, I introduced the classical planning problem and explained its relevance to the seminar by contrasting it with SAT. For those that haven't seen planning before, the aim was to provide some basic understanding of the problem and why it is of interest. For those familiar with planning, I attempted to give additional perspectives on the problem and its complexity. I aso gave the seminar participants a brief update on research in the planning community that tackles the main motivating questions of the seminar, in particular discussing results and open challenges for certifying planning algorithms.

3.15 Graph Symmetries, Patterns, and Encodings

Mikoláš Janota (Czech Technical University - Prague, CZ)

Joint work of Mikoláš Janota, Michael Codish
Main reference Michael Codish, Mikoláš Janota: "Breaking Symmetries with Involutions", in Proc. of the 31st International Conference on Principles and Practice of Constraint Programming, CP 2025, August 10-15, 2025, Glasgow, Scotland, LIPIcs, Vol. 340, pp. 8:1–8:17, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org//10.4230/LIPICS.CP.2025.8

When searching for graphs specifying certain conditions, the LexLeader approach is used to avoid symmetric graphs. A graph G is a LexLeader if its adjacency matrix is lexicographically smallest among all adjacent matrices representing graphs isomorphic to G. Symmetry breaking then amounts to adding constraints that eliminate non-LexLeader graphs from the search. We give a fresh perspective on symmetry breaking as a set cover problem and quantifier elimination problem [1]. A permutation π covers a graph G if G's adjacency matrix becomes smaller under π. We further define the notion of a *pattern* [2], which describe a set of graphs that become smaller because of some permutation π at position i. To encode patterns as CNF, extra variables are needed. These are Tseitin variables that describe an equality between two edge variables. Notably, these variables are reused across multiple patterns. Like so, a set of patterns enables us elegantly expressing a set of non-lexleaders and therefore can be interpreted as a certificate for a symmetry breaking constraint.

References

1 Michael Codish and Mikoláš Janota. *Breaking symmetries from a set-covering perspective.* In Proc. of the 22nd International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research, CPAIOR 2025, November 10–13, 2025, Melbourne, Victoria, Australia, LNCS, Vol. 15762, pp. 169–187, Springer, 2025.
2 Michael Codish and Mikoláš Janota. *Breaking symmetries with involutions.* In Proc. of the 31st International Conference on Principles and Practice of Constraint Programming, CP 2025, August 12-15, 2025, Glasgow, Scotland, LIPIcs, Vol. 340, pp. 8:1–8:17, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.

3.16 Certifying Ideal Membership Tests

Daniela Kaufmann (TU Wien, AT)

Deciding ideal membership is a central problem in computer algebra, with wide-ranging applications in geometry, verification, and symbolic computation. While Gröbner bases provide a complete method for deciding ideal membership, their outputs are often complex, making certification difficult.

I present a framework for certifying ideal membership tests using a practical algebraic calculus (PAC) that allows tracking polynomial manipulations. The calculus supports different levels of granularity, allowing proofs to be either concise or detailed; depending on whether the emphasis is on debugging or efficient proof checking.

3.17 Practically Feasible Proof Logging for Pseudo-Boolean Optimization

Wietze Koops (Lund University, SE & University of Copenhagen, DK)

Joint work of Wietze Koops, Daniel Le Berre, Magnus O. Myreen, Jakob Nordström, Andy Oertel, Yong Kiam Tan, Marc Vinyals
Main reference Wietze Koops, Daniel Le Berre, Magnus O. Myreen, Jakob Nordström, Andy Oertel, Yong Kiam Tan, Marc Vinyals: "Practically Feasible Proof Logging for Pseudo-Boolean Optimization", in Proc. of the 31st International Conference on Principles and Practice of Constraint Programming, CP 2025, August 10-15, 2025, Glasgow, Scotland, LIPIcs, Vol. 340, pp. 21:1–21:27, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org//10.4230/LIPICS.CP.2025.21

Certifying solvers have long been standard in Boolean satisfiability (SAT), allowing for proof logging and checking with limited overhead. However, developing similar tools for combinatorial optimization has remained a challenge. A recent promising approach covering a wide range of paradigms is pseudo-Boolean proof logging, but this has mostly consisted of proof-of-concept works far from delivering the performance required for real-world deployment.

In this work, we present an efficient toolchain based on VeriPB and CakePB for formally verified pseudo-Boolean optimization, and implement proof logging for the full range of techniques in the state-of-the-art solvers RoundingSat and Sat4j. Our experimental evaluation shows that proof logging and checking performance in this much more expressive paradigm is now quite close to the level of SAT solving, and hence clearly practically feasible.

3.18 Proof Logging for Subgraph-Finding Algorithms

Ciaran McCreesh (University of Glasgow, GB)

Many interesting problems involve finding a little graph inside a bigger graph: for example, maximum clique asks for the largest set of vertices where everything is adjacent, whilst subgraph isomorphism asks whether a specific pattern occurs inside a given target graph.

Although these problems are computationally hard in theory, in practice solvers can often handle these problems extremely quickly, even on graphs with thousands of vertices. However, these solvers are not always perfect, and sometimes contain bugs that lead to wrong answers being produced. I'll explain how, using the VeriPB proof system, we can augment these solvers to produce correctness certificates, allowing us to be confident they have definitely given the right answers. To do this, we'll need to be able to justify a wide range of algorithmic inference steps, including colour bounds, all-different filtering, and degree reasoning; perhaps surprisingly, VeriPB is able to do all of these efficiently, despite not having any notion of what a graph is.

3.19 Certified Constraint Programming

Matthew McIlree (University of Glasgow, GB)

Joint work of Matthew McIlree, Ciaran McCreesh, Stephan Gocht, Jakob Nordström, Jan Elffers
Main reference Stephan Gocht, Ciaran McCreesh, Jakob Nordström: "An Auditable Constraint Programming Solver", in Proc. of the 28th International Conference on Principles and Practice of Constraint Programming, CP 2022, July 31 to August 8, 2022, Haifa, Israel, LIPIcs, Vol. 235, pp. 25:1–25:18, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.
URL https://doi.org//10.4230/LIPICS.CP.2022.25

Constraint programming (CP) is a powerful paradigm for expressing and solving satisfaction and optimisation problems involving finite domain variables and high-level constraints. But the implementation and engineering of CP algorithms can be extremely complex, error-prone, and difficult to test. We are much more likely to trust the output of a solver if it can provide some kind of certificate of correctness via proof logging.

In this talk, I will discuss the current state of research into adding proof logging to CP solvers. I'll cover how we can prove unsatisfiability and optimality; what makes this different from established proof logging technology for SAT solvers; and the efforts towards devising efficient justification procedures for the huge variety of propagation algorithms available in the modern CP repertoire.

3.20 Certifying Presolving/Preprocessing for 0-1 Integer Linear Programming and MaxSAT

Andy Oertel (Lund University, SE)

Joint work of Jeremias Berg, Ambros Gleixner, Alexander Hoen, Hannes Ihalainen, Matti Järvisalo, Magnus Myreen, Jakob Nordström, Andy Oertel, Yong Kiam Tan
Main reference Alexander Hoen, Andy Oertel, Ambros M. Gleixner, Jakob Nordström: "Certifying MIP-Based Presolve Reductions for 0-1 Integer Linear Programs", in Proc. of the Integration of Constraint Programming, Artificial Intelligence, and Operations Research - 21st International Conference, CPAIOR 2024, Uppsala, Sweden, May 28-31, 2024, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 14742, pp. 310–328, Springer, 2024.
URL https://doi.org//10.1007/978-3-031-60597-0_20
Main reference Hannes Ihalainen, Andy Oertel, Yong Kiam Tan, Jeremias Berg, Matti Järvisalo, Magnus O. Myreen, Jakob Nordström: "Certified MaxSAT Preprocessing", in Proc. of the Automated Reasoning - 12th International Joint Conference, IJCAR 2024, Nancy, France, July 3-6, 2024, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 14739, pp. 396–418, Springer, 2024.
URL https://doi.org//10.1007/978-3-031-63498-7_24

It is well known that reformulating the original problem can be crucial for the performance of mixed-integer programming (MIP) and maximum satisfiability (MaxSAT) solvers. While the idea is the same in both the MIP and MaxSAT community, the presolving reductions in MIP and preprocessing in MaxSAT apply slightly different techniques to reformulate the problem. To ensure the correctness of the reformulations, all transformations must preserve the feasibility and optimal value of the problem, but there is currently no established methodology to express and verify the equivalence of two optimization problems.

In this talk, it is presented how pseudo-Boolean proof logging can be used to certify the correctness of a wide range of modern MIP presolving and MaxSAT preprocessing techniques. By combining and extending the VeriPB and CakePB tools, we obtain a formally verified end-to-end proof checking tool chain to verify the correctness of reformulations of pseudo-Boolean problems.

This talk is based on the following two papers. The first paper was published at CPAIOR 2024 together with Alexander Hoen, Ambros Gleixner, and Jakob Norström. The second paper was published at IJCAR 2024 together with Hannes Ihalainen, Yong Kiam Tan, Jeremias Berg, Matti Järvisalo, Magnus O. Myreen, and Jakob Nordström.

3.21 Certified First-Order Theorem Proving: confessions, excuses and a few ways out.

Michael Rawson (University of Southampton, GB)

Joint work of Michael Rawson, Anja Petković Komel, Martin Suda

Automated Theorem Provers (ATPs) have been around a long time, but their proof certification ecosystem is nowhere near as well-developed as, say, SAT solvers. There are several reasons for this: a plurality of proof calculi, equisatisfiable inferences, and theories, to name a few. I will outline ATP systems and their proof certification, explain some difficult areas, present some recent developments, and offer a few paths to salvation.

3.22 Trimming SMT Proofs

Joseph Reeves (Carnegie Mellon University - Pittsburgh, US)

Joint work of Joseph Reeves, Haniel Barbosa, Marijn J. H. Heule, Andrew Reynolds

Automated reasoning tools require high trust, prompting modern solvers to produce proof certificates for verification. In propositional satisfiability (SAT), proof generation and checking are relatively inexpensive, but in satisfiability modulo theories (SMT), justifying theory lemmas can introduce significant overhead. Recent approaches mitigate this issue by having the SMT solver produce a proof skeleton containing only the propositional reasoning in the DRAT format and unjustified theory lemmas, whose justifications are deferred to the checking phase. Preprocessing, a key element of SMT solving that can be challenging to justify a posteriori, is not justified nor checked. We extend these approaches by including proofs for preprocessing; by reducing the checker workload via iteratively eliminating theory lemmas from proof skeletons through SAT solving and proof trimming; and by proposing two justification methods for theory lemmas: one batches justifications for parallelization, while the other not only checks the theory lemma justifications but also integrates them into a fully detailed proof that could be checked with standard approaches. Experimental results on SMT-LIB benchmarks show the benefits of our approach in reducing solving time when producing proof skeletons that can be effectively checked externally. In particular, the extended trimming techniques can significantly reduce the number of theory lemmas to be checked beyond standard trimming, thereby improving sequential and parallel checking times.

3.23 Engineering Complete SMT Proofs in *cvc5* with Ethos/Eunoia

Andrew Reynolds (University of Iowa - Iowa City, US)

Over the past 5 years, the SMT solver *cvc5* has been instrumented to produce proofs for a majority of its theories. This talk reports on a new milestone for this work, namely that all mainstream features of *cvc5* are 100% proof producing and checkable in an external proof checker (Ethos). In detail, Ethos is a high performance proof checker written in around 10k lines of C++. Its native input language is Eunoia, a logical framework for defining proof systems that is heavily inspired by the forthcoming SMT-LIB version 3.0 language. To engineer complete proofs for *cvc5*, I will discuss the introduction of a "safe mode" of *cvc5*, which defines a subset of the features of *cvc5* that are free of known bugs and have complete proof support. The internal proof calculus of *cvc5*, now known as the Cooperating Proof Calculus (CPC), has been formalized in around 6500 lines of Eunoia definitions. Notably, this formalization now covers all mainstream theories of *cvc5*, including those currently used by industrial users of *cvc5*.

3.24 Pseudo-Boolean Proof Logging for Optimal Planning

Tanja Schindler (Universität Basel, CH)

Joint work of Simon Dold, Malte Helmert, Jakob Nordström, Gabriele Röger, Tanja Schindler
Main reference Simon Dold, Malte Helmert, Jakob Nordström, Gabriele Röger, Tanja Schindler: "Pseudo-Boolean Proof Logging for Optimal Classical Planning", in Proc. of the 35th International Conference on Automated Planning and Scheduling, ICAPS 2025, November 9-14, 2025, Melbourne, Victoria, Australia, Proc. ICAPS, Vol. 35(1), pp. 54–63, AAAI Press, 2025.
URL https://doi.org//10.1609/ICAPS.V35I1.36101

Optimal classical planning is the problem to find an action sequence with minimal cost from a given initial state to a goal state. Checking that a given action sequence is a plan can easily be done by applying the actions one after another, but if a planning system claims that a plan it has found is optimal or that there is no plan, a different kind of proof is needed. In my talk, I present our recent efforts to provide such a proof in the form of lower-bound certificates based on pseudo-Boolean constraints. These certificates can then be checked by VeriPB. I demonstrate how planning systems based on heuristic search can be modified to produce such certificates, and discuss the current status of the approach.

3.25 Certified QBF Solving

Martina Seidl (Johannes Kepler Universität Linz, AT)

Over the last years, much progress has been made in theory and practice of solving quantified Boolean formulas (QBFs). In principle, it is also well understood how to certify solving results found on solvers based on different solving paradigms like QCDCL as well as abstraction- and expansion-based solving. QBF certification is strongly inspired by approaches successfully used in SAT. Nevertheless, state-of-the-art solvers support certification to a limited extent only.

In this talk, an overview is given on the state of the art of certified QBF solving and the challenges that need to be addressed to obtain a fully operational certification workflow.

3.26 Certifying Algorithms in Railway Verification

Monika Seisenberger (Swansea University, GB)

Joint work of Harry Bryant, Alec Critten, Andrew Lawrence, Monika Seisenberger, Anton Setzer

We report on one old and some recent advances we made in the context of applying SAT/SMT solving in the area of Railway Verification.

The first concerns a provably correct DPLL/Resolution solver [1] that was extracted from a formal proof (in the theorem prover Minlog) that for every clause set there is either a satisfying assignment of variables or a proof of unsatisfyability. The extracted program from this (Minlog) proof yields a program (in Haskell) that either computes a model or a (DPLL system) proof of unsatisfiability. The extracted solver was applied to a number of Railway problems.

Our new work concerns a certified RUP checker that has been extracted from a formal proof in the Theorem Provers Rocq and Agda [2]. We formalised the RUPchecker in Rocq, provided a soundness proof and extracted the checker from it. The procedure also allows to produce the corresponding Unitresolution proof, but to do so is not required for the correctness of the checked result.

This is joint work with Harry Bryant, Alec Critten, Andrew Lawrence (Siemens Mobility), and Anton Setzer.

References

1 Ulrich Berger, Andrew Laurence, Fredrik Nordvall Forsberg, and Monika Seisenberger. *Extracting verified decision procedures: DPLL and Resolution*. Logical Methods in Computer Science, Vol. 11(1), 2015. https://doi.org//10.2168/LMCS-11(1:6)2015

2 Harry Bryant, Andrew Lawrence, Monika Seisenberger, Anton Setzer. *Verifying Z3 RUP proofs with the interactive theorem provers Coq/Rocq and Agda*. In Proc. of the 31st International Conference on Types for Proofs and Programs, TYPES 2025, June 9-13, 2025, Glasgow, Scotland, Abstracts, 2025.

3.27 Proof Verification with GDV and LambdaPi - It's a Matter of Trust

Geoff Sutcliffe (University of Miami, US)

Joint work of Geoff Sutcliffe, Frédéric Blanqui, Guillaume Burel
Main reference Geoff Sutcliffe, Frédéric Blanqui, Guillaume Burel: "Proof Verification with GDV and LambdaPi - It's a Matter of Trust", in Proc. of the 38th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2025, Daytona Beach, FL, USA, May 20-23, 2025, Florida Online Journals, 2025.
URL https://doi.org//10.32473/FLAIRS.38.1.138642

Automated Theorem Proving (ATP) is concerned with the development and use of software that automates sound reasoning. An ATP system can be required to output a proof that serves as a certificate for the system's claim. To ensure that a proof is correct, verification can be required. If the verifier outputs evidence in a form that can be independently checked, that evidence serves as a certificate for the verifier's claim. The sequence of finding a proof, verifying the proof, and certifying the verification, builds an increasing level of trust in the system. This talk traces one such path for TPTP format proofs generated by ATP systems, via the GDV derivation verifier, and ending at the LambdaPi checker.

References

1 Geoff Sutcliffe, Frédéric Blanqui, Guillaume Burel. *Proof Verification with GDV and LambdaPi - It's a Matter of Trust*. In Proc. of the 38th International FLAIRS Conference, FLAIRS-38, May 20-23, 2025, Daytona Beach, Florida, USA, Proc. FLAIRS, Vol. 38, Florida Online Journals, 2025.

3.28 Certifying Dynamic Symmetry Breaking for Graph Search in SAT and QBF

Stefan Szeider (TU Wien, AT)

Joint work of Mikoláš Janota, Markus Kirchweger, Tomás Peitl, Stefan Szeider
Main reference Mikoláš Janota, Markus Kirchweger, Tomás Peitl, Stefan Szeider: "Breaking Symmetries in Quantified Graph Search: A Comparative Study", in Proc. of the AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, pp. 11246–11254, AAAI Press, 2025.
URL https://doi.org//10.1609/AAAI.V39I11.33223

SAT Modulo Symmetries (SMS) performs dynamic symmetry breaking for graph generation by detecting and excluding non-canonical graphs during CDCL search. In this talk, we will focus on the certification mechanisms that ensure the correctness and completeness of this approach across different settings.

We will consider three types of certificates: (1) nc-certificates (non-canonicity certificates), which are permutations proving that a partially defined graph cannot be extended to a lexicographically minimal solution; (2) DRAT proofs where the symmetry-breaking clauses learned via the minimality check are added as axioms to the proof, with these axioms themselves certified by their corresponding nc-certificates; and (3) uniform proofs for the QBF setting, where SMS handles quantified graph search problems with formal verification through LDQ-resolution.

These certification mechanisms provide mathematical guarantees for the correctness of SMS and enable independent verification of results in challenging combinatorial problems, from confirming the Murty-Simon conjecture to computing Ramsey graphs with formal proofs.

3.29 The Past, Present, and Future of Verified Proof Checkers

*Yong Kiam Tan (Institute for Infocomm Research (I2R), A*STAR, Singapore, SG & Nanyang Technological University, Singapore, SG) and Magnus Myreen (Chalmers University of Technology - Göteborg, SE)*

URL https://cakeml.org/checkers.html

This survey talk will be split into two parts.

First, we will survey various automated reasoning theories/domains where verified proof checkers have been built. We will also present some of our ongoing work, and we will argue that verification can help enable the design of more complex proof systems/checkers while preserving trust in the overall certification process.

Then, we will take a deeper dive into verification infrastructure available in various theorem provers. Special focus will be given to *HOL4* and the *CakeML* project – we have used *CakeML* to build several end-to-end verified proof checkers with machine-code level correctness guarantees. We will discuss synergies between what *CakeML* can bring to proof checking and what proof checking can bring to *CakeML*.

3.30 Certification in SCL

Christoph Weidenbach (MPI für Informatik - Saarbrücken, DE)

SCL is a new paradigm for automatic reasoning in various logics. I discuss certification of proofs in SCL.

4 Panel Discussion: The Future of Certifying Algorithms

On Thursday afternoon, a 1-hour panel discussion with the title *The Future of Certifying Algorithms* was organized. The panelists were Marijn Heule, Haniel Barbosa, Ciaran McCreesh, and Ambros Gleixner, and the session was moderated by Jakob Nordström.

The panel started with an opening statement by each panelist, where they could choose to either address a (suitably provocative) question provided by the panel moderator, or to talk about some other topic of their own choice. After that, the audience was invited to ask questions. Below follow summaries of the opening statements and some of the ensuing discussion.

4.1 Opening Statements

Marijn Heule. *Question*: "Is the point of proof logging to create terabyte-size or petabyte-size proofs that nobody can understand? Is this the best way in which combinatorial solving can help mathematics, or can we go beyond this?"
Response: Proof logging for SAT has been successful despite the size of the proofs. Finding a proof is hard, but making the proofs more compact using trimming and checking such smaller proofs is easier. So trimming is key to the success of SAT proof logging. In this seminar there have not been enough talks about trimming. Proof trimming techniques should be higher on the agenda for proof logging efforts that go beyond SAT.

Another question is when we can produce small certificates. For example, for some maximum clique-finding problems, one can certify optimality just by giving a coloring. So there should be more work on finding small proofs.

Haniel Barbosa. *Question*: "Why would SMT proof logging have to be so slow? SAT solvers have blisteringly fast proof logging despite learning upwards of 50,000 lemmas per second – are we saying that SMT solvers are doing substantially more reasoning per time unit?"
Response: The main issue is that SMT solvers are doing 30 procedures that are each as hard as SAT solving. So far, the SMT community has focused on producing proofs that can be checked, and ensuring that each component of the solver can output justifications. The proof checking overhead is higher than it should be, but there is just too much to do with all the different components of the SMT solver.

Ciaran McCreesh. *Question*: "Designing proof logging for more and more applications is all well and fine, but what is the point if the *VeriPB* proof system is so complicated that nobody can use it? Should proof logging beyond SAT require a PhD in computational complexity theory?"

Response: It should not be socially acceptable to claim to have a faster solver without implementing proof logging. To do proof logging, people do need to have some understanding, but usually it is not necessary to understand all of *VeriPB*. But it does not have to be hard: proof logging for clique (which does not use the so-called redundance-based strengthening rule) was done by a master student. It is not that much extra that we are asking.

Ambros Gleixner. *Question*: "What could possibly make the operations research (OR) community care about certifying solvers? And even if they cared, would closed-source commercial solvers ever release their proof logs, or would they worry more about leakage of commercial solver secrets?"
Response: Customers are just interested in finding the best primal solution as fast as possible, and this is what drives commercial MIP solvers. The customer is happy with a good feasible solution, while optimality is mainly to give a criterion when the solver can stop looking for better solutions. It is not in the economic interest of commercial parties to implement proof logging. Customers are currently unaware of any problems. The question of whether we can do proof logging without revealing every secret is a good academic research question.

Operations research is not just about commercial MIP solvers, but also about solving practically relevant problems. To convince the OR community, we need to show examples where people are interested in certified correctness, e.g., kidney exchange problems, or combinatorial auctions for which there are legal fairness requirements. There is hope, but we can only do convince people if we can show that we can do proof logging, and we are not there yet. But we should identify practical problems and show that we can solve them with proof logging.

4.2 Discussion

The ensuing discussion touched upon a range of topics as outlined below.

SAT competition track for producing small proofs. Dirk Beyer proposed to have a SAT competition track where the goal is to produce small proofs, rather than necessarily solve the most instances. Marijn Heule agreed that this was an excellent idea, although it might be tricky to get the incentives right. A question is whether the evaluation metric should be the number of proof steps rather than the checking time.

Using proof logging as debugging tool. To facilitate the use of proof logging and checking as a debugging tool, it would be desirable to have an interactive proof checker, so that the proof could be fed to the checker line by line and it would be possible to stop the checker and study what the internal state is. Also, to facilitate such usage of proof logging, the proof format should easier to read by humans, for instance, by allowing general variable names (instead of just numbers as in *DRAT*).

To what extent is certification required? There was a discussion about the papers by Mehlhorn et al. [1, 2] on certifying algorithms. A provocative question: Should we teach in our undergraduate algorithm courses that every algorithm should be certifying? The general consensus was that this would be too extreme. While there are some nice examples of certifying algorithms, these are in the minority, and it is not really doable to have certification for every single algorithm. The proof logging may not even be related to how we justify the algorithm theoretically. One should not get away with not thinking about certification at all, but banning algorithms that are not certifying would be too radical.

Manifesto on proof logging. Next, there was a discussion that it would be nice to have a manifesto on proof logging. Such a document could explain in detail what is meant by a certifying algorithms, and what different flavours of proof logging and proof checking are employed in different context. It could also establish common terminology for different concepts that arise in the context of certifying algorithms. Finally, such a manifesto could be a valuable reference, that could be cited to substantiate that research on certifying algorithms is an important endeavour.

Outcomes of the seminar

Towards the end of the panel discussion, there was a conversation about the to what extent the seminar week had been successful. It was noted that the seminar had gathered researchers from many different communities working on certifying algorithms, and that it was interesting and useful to compare and contrast different proof logging approaches. Meeting and talking, and in this way creating a common understanding in between different communities, had been valuable, and developing a common language makes it possible to communicate, and collaborate, more productively going forward. Some seminar participants pointed out, in particular, that it had been very interesting to learn about techniques for building formally verified proof checkers as a way to provide combinatorial solving with end-to-end verification.

References

1 Eyad Alkassar, Sascha Böhme, Kurt Mehlhorn, Christine Rizkallah and Pascal Schweitzer. *An Introduction to Certifying Algorithms.* Information Technology Methoden und innovative Anwendungen der Informatik und Informationstechnik, Vol. 53(6), pp. 287–293, 2011.

2 Ross M. McConnell, Kurt Mehlhorn, Stefan Näher and Pascal Schweitzer. *Certifying Algorithms.* Computer Science Review, Vol. 5(2), pp. 119–161, 2011.

4.3 Closing Statements

Finally, each panelist was given the opportunity to give a short closing statement.

Ciaran McCreesh. Certification can do really something good and useful to make world a better place. Algorithms have a bad name currently. Instead, "algorithm" should become a word that people trust, like they trust, for example, bridges and elevators not to collapse.

Marijn Heule. The tools that we are developing should be used more widely. For instance, there should be potential for proof logging in the context of large language models (LLMs). Also, it would be good to develop tools to find small proofs and small counterexamples.

Ambros Gleixner. For some applications, there will be a demand for certifying algorithms, but not for every application. The first order of business, though is to develop the tools that will make proof logging possible.

Haniel Barbosa. Developers of other combinatorial solvers will not want to go through the immense pain that it has been to make *cvc5* proof producing. Proofs should have fewer details, to make it easier to produce the proof and reduce the overhead.

5 Evaluation of the Seminar by Participants

In addition to the traditional Dagstuhl evaluation after the seminar, the organizing committee also arranged for a separate evaluation which specific questions about different aspects of the seminar. Below follows a (brief and selective) summary of the answers collected in both evaluations.

In the Dagstuhl survey, the scientific quality of the seminar was ranked very highly, and the seminar also scored highly on the questions whether it inspired new ideas, led to insights from neighbouring fields or communities, and inspired new research.

In the organizers survey (which was filled in by 21 participants – a bit less than the 28 persons filling in the official Dagstuhl survey), the decision *not* to have an open problem session was mostly assessed as good or very good. A majority of respondents to the organizer poll agreed that the discussion panel that was organized on Thursday afternoon was very good. Regarding the amount of scheduled activities in the seminar program overall, there was an even split between votes for "about right" and "a bit too much." A large majority found the balance between longer survey talks and shorter contributed talks in the program to be good, but a noticeable minority found the number of survey talks to be a bit too high.

Since the seminar tried to cover a fairly large number of different research areas related to automated reasoning and combinatorial solving, the organizer poll asked the participant whether there was a good balance between depth and breadth. Several participants commented that the coverage of many different areas was a good aspect of the seminar, and there were even suggestions for other areas to cover, such as knowledge representation and reasoning. Having more participants from industry (or more applied research) would also have been good, according to several respondents.

Among good aspects to keep for future editions in this seminar series the responses listed:

- The good balance of participants, including the mix of senior and junior researchers and coverage of different research areas.
- The balance between talks and informal discussions.
- The survey talks (where it would be good to think about how to make sure that there would be sufficient time for questions and discussions).
- The panel discussion.

Some aspects that could be improved were as follows:

- Maybe a "reading list" or similar could be provided to help participants prepare for the seminar (and to avoid survey talks having to start with the basics). Also, maybe some of the more "basic" survey talks could be scheduled in parallel?
- It would be good to get people to talk more about the problems they are encountering right now to encourage more future-looking discussions. For presentations of successfully concluded projects, there could be more of an emphasis on highlighting tools and techniques that could be useful also for other applications.
- It would be good to consider having an open problem session early on during the week, with an update on the last day on any progress made during the week.
- For future seminars on certifying algorithms, it would be good to to have fewer talks. It could also make sense to have working groups on different topics scheduled on short notice during the seminar week.
- More in-depth technical discussions of different proof formats, what they can do or not do, and what the pros and cons are of different approaches.
- Some hands-on demos of different proof logging tools would have been good.

All in all, it seems fair to say that the feedback from the participants was overwhelmingly positive. When asked if they would come to a similar seminar again in Europe, 20 out of 21 respondents in the organizer poll replied that this is very likely. If such a seminar were instead to be held in North America, a clear majority would still want to come, but the enthusiasm in the responses went down slightly. For a seminar in East Asia or India, the responses were even more mixed, with positive and negative votes perfectly balanced.

Participants

- Erika Ábrahám
RWTH Aachen University, DE
- Markus Anders
RPTU Kaiserslautern-Landau, DE
- Franz Baader
TU Dresden, DE
- Haniel Barbosa
Federal University of Minas Gerais-Belo Horizonte, BR
- Jeremias Berg
University of Helsinki, FI
- Dirk Beyer
LMU München, DE
- Armin Biere
Universität Freiburg, DE
- Nikolaj S. Bjørner
Microsoft – Redmond, US
- Bart Bogaerts
KU Leuven, BE
- Benjamin Bogø
University of Copenhagen, DK
- Randal E. Bryant
Carnegie Mellon University – Pittsburgh, US
- Sam Buss
University of California – San Diego, US
- Simon Dold
Universität Basel, CH
- Bruno Dutertre
Amazon Web Services – Santa Clara, US
- Katalin Fazekas
TU Wien, AT
- Mathias Fleury
Universität Freiburg, DE
- Ambros Gleixner
HTW – Berlin, DE
- Arie Gurfinkel
University of Waterloo, CA
- Malte Helmert
Universität Basel, CH
- Marijn J. H. Heule
Carnegie Mellon University – Pittsburgh, US
- Matti Järvisalo
University of Helsinki, FI
- Mikoláš Janota
Czech Technical University – Prague, CZ
- Daniela Kaufmann
TU Wien, AT
- Wietze Koops
Lund University, SE & University of Copenhagen, DK
- Konstantin Korovin
University of Manchester, GB
- Hanna Lachnitt
Stanford University, US
- Ciaran McCreesh
University of Glasgow, GB
- Matthew McIlree
University of Glasgow, GB
- Magnus Myreen
Chalmers University of Technology – Göteborg, SE
- Jakob Nordström
University of Copenhagen, DK & Lund University, SE
- Andy Oertel
Lund University, SE
- Albert Oliveras
UPC Barcelona Tech, ES
- Michael Rawson
University of Southampton, GB
- Joseph Reeves
Carnegie Mellon University – Pittsburgh, US
- Andrew Reynolds
University of Iowa – Iowa City, US
- Tanja Schindler
Universität Basel, CH
- Martina Seidl
Johannes Kepler Universität Linz, AT
- Monika Seisenberger
Swansea University, GB
- Mate Soos
Ethereum – Berlin, DE
- Geoff Sutcliffe
University of Miami, US
- Stefan Szeider
TU Wien, AT
- Yong Kiam Tan
Institute for Infocomm Research (I2R), A*STAR, Singapore, SG & Nanyang Technological University, Singapore, SG
- Dieter Vandesande
VU – Brussels, BE
- Christoph Weidenbach
MPI für Informatik – Saarbrücken, DE

Report from Dagstuhl Seminar 25232

Navigating the Maze of Guidelines to Unify Visualization Design Recommendations

Miriah Meyer*[1], Ghulam Jilani Quadri*[2], and Paul Rosen*[3]

1 **Linköping University, SE.** miriah.meyer@liu.se
2 **University of Oklahoma – Norman, US.** quadri@ou.edu
3 **University of Utah – Salt Lake City, US.** paul.a.rosen@gmail.com

Abstract

The field of visualization suffers from a persistent problem: guidance for visualization design is abundant but fragmented, unevenly evidenced, difficult to generalize across contexts, and often hard to access or teach. Further, these guidelines come from diverse sources, including theoretical foundations, empirical studies, design studies, and practitioner expertise. However, turning this knowledge into actionable forms of best practice remains an open problem. The goal of this seminar was to examine how guidelines are produced, interpreted, and operationalized, especially under pressures from domain specificity, communication stakes (e.g., misinformation and decision support), and the emerging role of generative AI in visualization workflows. The seminar challenged assumptions about the validity, transferability, and values encoded in guidelines through working groups on AI and guidelines, characterizing guidelines, values and teaching, and the goals for effective guidance.

Seminar June 1–6, 2025 – https://www.dagstuhl.de/25232
2012 ACM Subject Classification Human-centered computing → Empirical studies in visualization; Human-centered computing → Visualization; Human-centered computing → Visualization design and evaluation methods; Human-centered computing → Visualization theory, concepts and paradigms
Keywords and phrases design studies, qualitative evaluation, visualization design, visualization recommendations, visualization system and generative ai
Digital Object Identifier 10.4230/DagRep.15.6.32

1 Executive Summary

Paul Rosen (University of Utah – Salt Lake City, US)
Miriah Meyer (Linköping University, SE)
Ghulam Jilani Quadri (University of Oklahoma – Norman, US)

The field of visualization suffers from several interrelated challenges around design guidelines. First, we generate many loosely connected artifacts–theoretical frameworks, controlled experiments, qualitative studies, design studies, and practitioner expertise, etc. Second, there are challenges with generalization and the synthesis of research with little to no common framework that connects them (i.e., there is no good "theory of visualization"). Third, the artifacts we produce are hard to access – we produce many difficult-to-read papers, not to mention issues of education and literacy, communication and misinformation, role in decision making, etc.

* Editor / Organizer

Navigating the Maze of Guidelines to Unify Visualization Design Recommendations, *Dagstuhl Reports*, Vol. 15, Issue 6, pp. 32–50
Editors: Miriah Meyer, Ghulam Jilani Quadri, and Paul Rosen
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

At this seminar, we explored these challenges through the question: How do we formulate and integrate the knowledge we produce to best serve the visualization community and the world broadly?

The seminar started with lightning talks from all participants. The participants chose from a variety of provocations provided by the organizers to guide their talk content. The provocations were:

1. "Are Guidelines Just Bullsh*t?" What if the guidelines we cling to are nothing more than overgeneralized, useless lab artifacts?
2. "Shall Our AI Overlords Just Gobble Up Your Guidelines?" With AI generating visualizations autonomously, where do human-centered design principles fit in – or do they at all?
3. "Born in the Lab, Broken in the Wild?" Do our visualization guidelines reflect real-world needs, or just controlled experiments?
4. "Guidelines or Guardrails?" Are design guidelines empowering creativity, or are they limiting innovation with false certainty?
5. "Is Visualization a Science or a Craft?" If we treat vis as a scientific discipline, can we really generalize design? Or are we ignoring its artistic and contextual roots?
6. "Implications $\neq$ Instructions" Why do we keep mistaking exploratory study findings for universal design truths?
7. "Generalization Is a Comfort, Not a Guarantee" In our rush to codify design, are we sacrificing nuance and context for the illusion of control?
8. "Whose guidelines are these anyway?" Cognitive efficiency and perceptual accuracy underlie most visualization guidelines – is this all that we are about?

During the lightning talks, participants were encouraged to record ideas and thoughts on post-it notes, which the organizers used to create a set of themes for possible working groups. The entire group discussed the themes and agreed on the following ideas for working groups:

- AI + guidelines
- Characterizing guidelines
- Values + Teaching
- Goals for effective guidance

Working groups were encouraged to develop a zine by the end of the seminar to capture and communicate their main ideas. Three of the four groups produced zines; the fourth group created a short report document. One working group produced a panel proposal for the main visualization conference (IEEE VIS) as part of their working group. This panel was accepted and successfully run at the conference in November 2025.

To support cross-talk during the week, we had a mixer activity that took alternative themes for the seminar and had participants create a playlist of ideas for that theme. This activity resulted in new ideas feeding back into the existing working groups that broadened the scope of conversations.

The seminar resulted in a range of ideas, from concrete formulations of what exactly a guideline is and what makes it effective, to speculative ideas about the uses of generative AI for working with guidelines, and more far-reaching ideas about what values guidelines imply and what that says about the field of visualization more holistically.

2 Table of Contents

3 Overview of Talks

3.1 Data Visualization: Science, Craft, Both?

Bon Adriel Aseniero (AUTODESK – Toronto, CA)

This lightning talk explores the intersection of data visualization as both science and craft, drawing from my experience in navigating the design of visualizations in practice. Emphasizing themes that balances between structure and creative expression, this talk provokes reflection on the dual nature of visualization – as a science, anchored in empirical research, perception, and statistical integrity, and as a craft, deeply expressive, intuitive, and personal. The talk challenges the notion of fixed rules in visualization, emphasizing instead the diversity of ways a message can be visually conveyed. Like language, visualization has grammar and structure – but also metaphor, rhythm, and poetry. Rule-breaking can be useful when grounded in thoughtful design. Thus, I advocate for learning from vis designers' processes in creating visually compelling visualizations. The goal is not to prescribe, but to inspire dialogue and design grounded in human insight.

3.2 Questions for AI in visualizations

Michael Aupetit (HBKU – Doha, QA)

I mostly worked on data representation from geometrical encoding to visual perception of patterns. When data are transformed, encoded, and then decoded and interpreted by the end user, some information is lost.

Design guidelines can help at every step of the visualization pipeline, ensuring that as much information as possible is retained. It depends on the task we want to solve, the user, and the context in which it is used.

Generative Artificial Intelligence holds great promise in integrating a vast amount of human knowledge across all fields to provide guidance. However, they require a carefully designed loss function based on expert knowledge and a large amount of clean data to learn from them what constitutes good or bad guidelines for a prompted request.

The challenges are that the existing guidelines are limited to a few isolated cases from very different domains and abstraction levels. How do we scale up the number of cases and their variety while maintaining the data quality and meaningfulness? How can we encode such a diverse range of guidelines at various levels of abstraction? Is a language model combined with a vision model enough? Shall we use agentic approaches? Which roles for the agents? Do we have computational models that, in theory, can generate such guidelines? How do we validate such AI models?

3.3 Born in the Lab, Broken in the Wild?

Cindy Xiong Bearfield (Georgia Institute of Technology – Atlanta, US)

Why are we still doing lab studies?! This critique often emerges when lab-derived guidelines break down in messy, real-world settings. But perhaps the problem isn't the lab itself. It's how we frame its purpose and what we expect from it.

I propose that the lab serves two critical roles. First, it helps us understand why something works (or doesn't) through controlled studies. These studies isolate causal mechanisms, which can support generalization across formats and contexts. Success here means having explained something.

Second, the lab excels at rapid iteration. Testing and refining prototypes to make messy problems manageable. In this mode, success is measured by progress, not control. These studies are valid so long as we don't mistake them for universal truths.

Bad lab studies, then, fail in one of two ways: they stop too early (after a single surprising result), or they ask the wrong questions (focusing on trivial manipulations or ignoring essential context). I argue that we should design lab studies with the wild in mind. Instead of pitting lab against field, we should ask: What part of the wild are we trying to bring into the lab, and why?

3.4 How visualization design guidelines and visualization research relate

Fabian Beck (Universität Bamberg, DE)

The relationship between visualization design guidelines and visualization research is multifaceted and can be viewed from both research and practitioner perspectives. From a research standpoint, guidelines for designing visualizations may be perceived as vague, overly general, or even conflicting. They often lack the precision and rigor expected in scientific inquiry. However, from a practitioner's perspective, such guidelines serve as valuable tools for translating experience and domain knowledge into actionable recommendations for design practice.

Integrating these two perspectives is therefore highly relevant, though the nature of their relationship is not immediately clear. This relationship can be bidirectional: on one hand, research can lead to the formulation of new guidelines based on empirical findings; on the other hand, research can also investigate the implementation and effectiveness of existing guidelines in practical settings. Thus, the interplay between visualization research and design guidelines involves both the generation of guidance through scientific methods and the study of how such guidance is applied in real-world contexts.

3.5 The need for normative guidelines

Alexander Bock (Linköping University, SE)

Guidelines are a broad and sometimes diffuse concept while trying to span a multitude of different and diffuse use-cases and contexts. In my provocation I am arguing for the cases that this wide span of uses causes these codified guidelines to be either too rigid in their implementation, thus making them less useful in concrete application cases, or too broad to the level of being non-actionable. In particular with regard to the use of utilizing visualization to present to a diverse group of users simultaneously, we are currently limited to finding ad-hoc solutions that have to be rediscovered from first principles in every situation. So while there is a definite need for normative guidance or guidelines, the concrete description and application of such remains an open question to be solved by the community.

3.6 A case against generalization

Angelos Chatzimparmpas (Utrecht University, NL)

We like to believe that design rules (e.g., "bar charts are better than pie charts," or "use blue for low values and red for high") can be universally applied. However, in practice, context matters. What works well in a lab study with 100 student participants might fail in the wild with a different audience, task, or culture. So the comfort of generalization doesn't always hold in messy real-world settings. When we rely too heavily on codified design rules, we risk ignoring specific user needs, domain knowledge, and the aesthetic or emotional impact of design choices. This can stifle creativity and result in one-size-fits-all solutions that don't actually fit anyone well. Codified rules give designers the feeling of certainty, but that can be misleading. Following the rules doesn't necessarily mean we've made a good design decision. In most cases, it just means we've followed a "script".

In this lightning talk, I challenged the notion that pie charts, 3D visualizations, and rainbow colormaps are inherently "evil" by highlighting examples (based on academic papers) where, given the right context and audience, they can be effective. I also introduced one possible approach to addressing this debate: building a database that showcases various visualizations applied to the same data, aiming to bridge the gap in aligning design choices with human needs (in the spirit of tools like GraphScape). To further narrow the design space, I proposed simulating human perception using computer vision models. For instance, applying paradigms like the Just-Noticeable Difference (JND) could help evaluate whether a given visualization preserves enough perceptual signal to remain effective when substituted with an alternative.

I ended the talk with a provocation: what if the goal of visualization isn't to "express data effectively" but to maximize user engagement, spark curiosity, and invite diverse perspectives through interaction?

3.7 Guidelines: Right or Wrong, We Need Them

Michael Gleicher (University of Wisconsin-Madison, US)

Guidelines can help creators make less bad visualization (or make bad visualizations less often). However, to be effective, guidelines must have several properties. Correctness, that guidelines lead creators to the best designs, is only one property, albeit one that the academic community focuses on. However, I argue that properties relating to the actionability of guidelines – usability, credibility, and discoverability – are also important, maybe even more than correctness. This suggests that we need to develop an art/science of crafting guidelines, which has not been considered by the visualization community.

3.8 Is Visualization a Science or a Craft?

Lane T Harrison (Worcester Polytechnic Institute, US)

Visualization purports to be a science. It investigates perceptual and cognitive dimensions of visualizations, it develops visualization techniques with formalisms, it systematically studies the visualization design process and resulting artifacts. But visualization is a craft. Outside the walls of research, practitioners create visualizations with specific tools, develop skills, and apply these in a labor/professional context. Visualizations are created within specific contexts and settings. Visualizations are also created en masse, at frequencies and scales far larger than which they are studied. It would be insufficient to apply the current epistemologies and methodologies of visualization research to studying visualization as a craft. We need new approaches that allow us to study visualization as it actually is: collective, community-based, and culturally-situated.

3.9 What might visualization guardrails be?

Petra Isenberg (INRIA Saclay – Orsay, FR)

What we mean by (design) guidelines in the visualization community is less than clear. Guidelines can, on the one hand, be considered loose sets of rules that are meant to make things (processes/designs/systems/tools) effective in most cases. We should perhaps name such guidelines "considerations" – something people should consider and think about but that can certainly be broken when there is a good reason to. Such considerations have the problem that they require careful application, and therefore time and thought. They also require some form of empirical and practical backing to show when and where they applied well in the past or have been successfully broken. Guidelines might be, on the other hand, more easily applied if they were only framed at the level of guardrails – rules that, if broken, will likely lead to your process/design/system/tool to fail miserably. Yet, do we have any non-trivial or obvious guardrails in the community? How do we establish new ones? Are there ever rules that cannot or should not be broken?

3.10 Generalization about Visualization as a Decision Aid

Alex Kale (University of Chicago, US)

Visualization research often makes general claims about the usefulness of visualization as a decision aid. However, reflecting on the field's logic of generalization, I argue that we lack adequate ways of conceptualizing decision context and codifying the role it should play in design recommendations about decision aids. My talk sketches how decision theory provides a helpful framework for reasoning about the dimensions of decision context, identifying utility as a key but under-utilized way of accounting for how values and relationships around data shape the purpose of decision aids in practice.

3.11 Toward Human-centered Design Guidelines in the LLM – Revisiting Existing Guidelines

Sungahn Ko (POSTECH – Pohang, KR)

Traditionally, people have relied on textbooks and, more recently, the internet to access information. With the rapid rise in popularity of large language models (LLMs), driven by their impressive performance across various domains, it's increasingly likely that users will turn to LLMs for visualization-related tasks as well. For instance, LLMs can be used to generate new visualizations or assist with evaluating visualizations during data analysis workflows.

However, this introduces a concern: users without a background in visualization may struggle to assess whether the visualizations provided by LLMs are appropriate or effective. To address this issue, we need to explore the capabilities of LLMs in visualization-related tasks and propose evaluation criteria and guidelines to assess the quality and reliability of their visualization outputs.

3.12 AI meets visualization guidelines

Kuno Kurzhals (Universität Stuttgart, DE)

We identified four steps for the application of AI models in visualization design: (1) prompting, (2) style modification, (3) data-sensitive, and (4) user-sensitive modelling. While the first to types mainly serve the purpose of inspiration and prettying of visualizations, data-sensitive and user-sensitive models can explicitly incorporate guidelines for design on different levels such as appropriate mapping, aesthetics, and user-specific adaptation of the data representation. A combination of these steps can potentially solve a multitude of everyday visualization problems, without the requirement of visualization expertise to design a technique. However, the resulting visualizations are, like created by a designer, interpretations of a model and

require reasoning and a way to adjust the visualization, either by small or large changes. One big challenge of the near future will be how this communication between human user and model will look like. Only prompt-based discussions might be not efficient to discuss visual content that could be easier achieved by direct interaction with the visualized result.

3.13 Who's values are these, anyway?

Miriah Meyer (Linköping University, SE)

Looking across seminal visualization writings reveals a set of assumptions and values about what makes a visualization good. These values are: that the visualization is objective; people are universally preceptive, attentive, and predictable; and encodings are efficient. But these core values – that come from authors linked to the fields of stats and vision science – do not capture the breadth and diversity of visualizations today. They exclude emerging genres of visualizations for self-expression and rhetoric. They define a narrow range of what gets to count as a visualization, and who gets to make them. And they limit the influence of many early visualization pioneers who created visualizations from other perspectives with other goals. What other values can we align ourselves with to broaden the space of what counts as good?

3.14 Beyond Guidelines: Cultivating Visual Intuition

Carolina Nobre (University of Toronto, CA)

Most data visualization education follows a constraints-first approach: teach guidelines and theory, then provide opportunities for application through projects. My talk argues that leading with constraints limits creative exploration before students discover their visual voice. Drawing on evidence from teaching data visualization through hands-on "feel and see the data" exercises rather than theory-first instruction, the presentation proposes an alternate way of developing visualization literacy. When students build visual intuition first–unfettered by rules–they create compelling visualizations that communicate intent effectively, even when breaking conventional guidelines. This provocation asks whether we are teaching visualization literacy or visual compliance, and when design guidelines become barriers to developing authentic visual intelligence.

3.15 Lost in Translation: How and Who Should Be Applying These Guidelines?

Ghulam Jilani Quadri (University of Oklahoma – Norman, US)

In the rapidly evolving field of information visualization, rigorous evaluation is essential for validating new techniques, understanding user interactions, and demonstrating the effectiveness and usability of visualizations. These empirical studies yield practical, insightful, and innovative design guidelines for creating compelling and expressive visualizations, as well as their design choices. However, many times these guidelines are isolated, less connected, and challenging to bring together and combine implementation, leading to a situation of "Empirical Explosion – Practical Paralysis". In this talk, I introduced the empirical explosion based on prior surveys and predicted trends in historical data to showcase how an increasing number of empirical studies might advance the visualization community, and how they are not easily applied in the real world – the provocation questions aimed to spark interesting discussions.

3.16 So. . . why are we here?

Paul Rosen (University of Utah – Salt Lake City, US)

In this talk, I outline the origin of this Dagstuhl Seminar, namely, that the field of visualization suffers from several interrelated challenges around design guidelines. First, we generate many loosely connected artifacts – theoretical frameworks, controlled experiments, qualitative studies, design studies, and practitioner expertise, etc. Second, there are challenges with generalization and the synthesis of research with little to no common framework that connects them (i.e., there is no good "theory of visualization"). Third, the artifacts we produce are hard to access–we produce many difficult-to-read papers, not to mention issues of education and literacy, communication and misinformation, role in decision making, etc. Finally, I highlight a call to action – how do we formulate and integrate the knowledge we produce to best serve the visualization community and the world broadly?

3.17 Who's Guidelines Are These Anyways?: Beyond Cognitive Efficacy and Perceptual Accuracy

Arvind Satyanarayan (MIT – Cambridge, US)

Researchers and designers tend to focus the bulk of their effort on the accurate and efficient transmission of objective insights about data. Thus, when looking to diagnose failures in data communication, we look to intervene at steps along this transmission process: improving

encoding (e.g., through better design guidelines) or decoding (e.g., by boosting data/visualization literacy). However, guidelines that are primarily concerned with encoding-decoding cannot account for the full range of behaviors we have recently witnessed – particularly around how visualizations can propagate misinformation. This provocation suggests we need to look beyond the encoding-decoding model. By drawing on sociolinguistics, this provocation suggests we need to study "social inferences": the meanings people read into visualizations that are not about the data, but rather are about the identities and characteristics of the visualization author, and about their relationship to the reader. Initial evidence indicates that such social inferences mediate how receptive readers are to the information a visualization depicts, and how likely they are to further engage with the visualization.

3.18 Why do we keep mistaking exploratory study findings for universal design truths?

Karen Schloss (University of Wisconsin – Madison, US)

People likely mistake exploratory study findings for universal design truths because they want to be told what to do to guarantee "good" design without having to think critically about design (either because they do not have the tools/knowledge/intuitions or they do not have time/motivation to put in the effort). Through this lens, critical questions arise: what are concerns about assuming study implications do equate to instructions of how to produce "good" design, and what might we do about those concerns? My primary concern is that treating laboratory findings as "rules", in service of pithy, impactful guidelines can lead to two key problems. First, if people overgeneralize and use those rules where they do not apply, that can lead to problematic design choices. Second, if they ignore those rules because they think the rules don't apply in their context, that can lead to designers feeling guilt, anxiety, and even shame for "breaking" rules. Yet, there are opportunities to leverage lab findings to inform design, while accounting for nuance. One approach is through tools with adjustable parameters that enable designers to weight distinct design priorities without strong constraints. Still, more knowledge is needed to produce comprehensive, actionable tools. A potential frame for moving forward is to think in terms of the goals of the designer and the tasks of the observers, and what design properties support those goals. In doing so, it is important to think in terms of abstraction, developing theories/principles that will enable generalization of laboratory findings, rather than one-off guidelines. This approach calls evaluation tools, such as linters, into question – if linters are built on incomplete knowledge about how people interpret visualizations, that can lead to flagging effective designs as "wrong", passing problematic designs as "correct", and becoming overconfident that a design is effective having been checked and judged as "correct". Going forward, a key challenge is understanding how to convey nuanced lab findings in ways that are easy to understand and actionable, despite incomplete knowledge.

3.19 A case for looking forward

Michael Sedlmair (Universität Stuttgart, DE)

In visualization research, a lot of focus is placed on established design guidelines. While some of these can be effectively and illustratively communicated, they can also lead to overly authoritative generalizations – such as the belief that 3D pie charts and rainbow colormaps are always bad. In reality, the appropriateness of such choices often depends on context, and in many cases, they may not be as harmful as we assume – they are what they are: simply guidelines, not strict rules. Nevertheless, the community continues to invest considerable time in discussing and refining long-established rules and loses additional time by rejecting novel work for not strictly adhering to these easy-to-check "rules".

In my opinion, this time could be better spent identifying the next grand challenges the field is likely to face. These challenges could provide direction for the community and support the formation of subgroups focused on driving specific areas forward. One example of a forward-looking area we're exploring in our lab is situated visualization in augmented reality (AR). AR has the potential to become a fundamentally new medium for interacting with data and digital content. This technological shift could profoundly impact how we engage with data – much like mobile devices reshaped computing and communication over the past two decades.

To summarize, my provocation is this: we are currently spending too much time re-evaluating old guidelines that, in many cases, are already "good enough." Instead, I argue for a more future-oriented approach – developing design guidelines, considerations, patterns, and recommendations that address the new and emerging challenges ahead.

3.20 Guidelines developed in the lab versus in the wild

Vidya Setlur (Tableau Research – Palo Alto, US)

My talk challenges the disconnect between visual analytics design guidelines developed under controlled, idealized conditions and the complex realities of how people interact with data in practice. While existing guidelines emphasize curated dashboards and data, well-formed natural language queries, and visual minimalism, the real-world use is far messier; users are frequently under time pressure, accessing data insights on mobile devices, or multitasking across contexts. Their questions are often vague, exploratory, or evolving, and the data itself may be incomplete, inconsistently labeled, or lacking the metadata needed to support traditional interface assumptions.

This gap between designed expectation and actual experience demands a fundamental rethink of our guidelines. We need to center the realities of ambiguity, intent uncertainty, and user preference if we want our tools to support meaningful sensemaking – not just in the lab, but in the wild. This talk presents three provocations to reframe how we approach the design of visual analytics tools:

Provocation 1: Dashboards are often celebrated for their polish, well-thought out layouts, rich interactivity, and text – all of which break down in dynamic, mobile, or high-pressure environments. We must move beyond guidelines merely optimized for perfectly curated dashboards, and toward design principles that accommodate real-world conditions – time-constrained and on-the-go.

Provocation 2: Natural language interfaces for data assume clean semantics and well structured queries. We need to move beyond natural language interaction guidelines that expect clarity and completeness. Real users bring ambiguity, uncertainty, and shifting goals, and our systems must be designed to meet them there.

Provocation 3: Despite design dogma around minimalism and reduced data-ink, users often prefer densely annotated charts, descriptive captions, or even pure text summaries. We must move beyond design guidelines that prioritize visual minimalism, and acknowledge user preferences and personalization for rich explanatory context – through annotations, captions, or even standalone text.

3.21 Meaningfully specific, pluralistically rich – rethinking evidence and focus for visualisation design guidelines

Cagatay Turkay (University of Warwick – Coventry, GB)

(written in response to the provocation: "Generalisation is a comfort, not a guarantee")

I argue that we need to think about the form and the evidence basis of guidelines to bring nuance and context in a rich and informative way. Visualisation will benefit from a pluralistic approach to knowledge and ways of knowing and by valuing diverse forms of evidence to construct guidelines. We can look up to fields such as medicine and health sciences and learn from their practice in evidence synthesis but also learn from their mistakes – such as creating a hierarchy within methods and knowledge that has stifled diversity and richness of information. I would like to also argue that we also need to focus on nuance and context that matters in the world. I would like to see us developing guidelines for substantial issues, for instance, what are some guidelines to communicate climate change, social inequalities or quality of information – we need to be specific with guidelines but we need meaningful specificity.

3.22 Better visualization with Guidelines for/by AI or humans?

Tatiana von Landesberger (Universität Köln, DE)

Joint work of L. Pelchmann, L. Theile, S. Pandey, Team VisVA; M. Pohl, K. Ballweg, M. Wallner, Team TU Darmstadt

Our research group conducts research in three areas:

1. basic research on network, time series and geographic data perception and visualization with lab studies that derive guidelines for better data visualization based on lab experiments.

2. development of novel visualization techniques, where the basic principles of visualization guidelines are used and extended with novel visual designs tested in lab or crowdsourced experiments.
3. development of visual analytics systems that apply the guidelines from basic research and visualization techniques for addressing specific application needs. That leads to best practices and general guidelines for future general visual designs.

Visualization and guidelines have 4 different forms, two and two are human-applied or computer applied.

1. Human applied:
   a. practitioners create visualization based on their knowledge. This may or may not be based or use guidelines, rather their practical experience. They use a lot of context and domain information for development of visualizations.
   b. visualization experts: either researchers or experts in visualization use both their knowledge of guidelines and domain. The guidelines may not be formal.
2. computer applied
   a. Computer uses a collection of formal guidelines that it mechanically applies based on the guideline application criteria (see paper reference below). No context is taken into account.
   b. AI that creates visualizations. It is unclear whether and how much AI adheres / uses the existing guidelines, context and other criteria for good visualization.

3.23 Visualization: Science or Engineering?

Daniel Weiskopf (Universität Stuttgart, DE)

For this lightning talk, I slightly modify the provocation given by the Dagstuhl Seminar organizers ("Is Visualization a Science or a Craft?") to "Visualization: Science or Engineering?" Following Brooks' statement, "the scientist builds in order to study; the engineer studies in order to build" [F. Brooks, Comm. ACM 39(3), 1996], I see the visualization research community as a whole primarily as an engineering discipline. However, we draw a lot of methods, approaches, and findings from disciplines including but not limited to computer science, psychology, mathematics, social sciences, life sciences, art, and design. This breadth has an impact on how we deal with research questions centered around guidelines. One proposed consideration is to look into guidelines on the process rather than the result (i.e., visualization artifact), broadening the perspective on guidelines to make them more flexible. Another consideration is to articulate the scope, limits, and uncertainty of guidelines, thus addressing the possible issue of over-generalization. Finally, not one size fits all, i.e., there is a need for adequate guidelines, recommendations, or examples that may span, e.g., from low-level color perception all the way to a comprehensive visual analytics system. I also touch on a few other aspects that could be discussed during the Dagstuhl Seminar.

4 Working groups

4.1 GUIDELINES ARE NOT RULES: Characterizing Terminologies around Datavis Design Guidelines

Bon Adriel Aseniero (AUTODESK – Toronto, CA), Cindy Xiong Bearfield (Georgia Institute of Technology – Atlanta, US), Petra Isenberg (INRIA Saclay – Orsay, FR), Ghulam Jilani Quadri (University of Oklahoma – Norman, US), Paul Rosen (University of Utah – Salt Lake City, US), Karen Schloss (University of Wisconsin – Madison, US), and Daniel Weiskopf (Universität Stuttgart, DE)

We created a zine to summarize the ideas from our working group.

4.2 The 3Ps of Effective Guidance: Properties, Packaging, Process

Michael Gleicher (University of Wisconsin-Madison, US), Michael Sedlmair (Universität Stuttgart, DE), and Cagatay Turkay (University of Warwick – Coventry, GB)

We want to create "guidelines" that help designers make better visualizations. To do this, we need to understand the properties that guidelines should have in order for them to be effective at this goal. Currently, there is very little guidance on how to create and evaluate "good" design guidelines (meta guidance). We need to (1) identify and characterize the properties of good guidelines, (2) devise ways to package them, and then (3) describe how to embed them into the design process.

We followed the following steps in our working group:

Brainstorming properties: Our process began with brainstorming an initial set of properties and identifying other resources that provide guidance on guidelines. In particular, we looked to the medical domain where they have formalized evaluation criteria for guidelines development [1, 2]. This led to a large list of 30 initial properties. We grouped them into 11 themes using affinity diagramming.

Formative evaluation and iteration: Using the 11 themes, we did a cognitive walkthrough, in which we systematically compared our properties to those listed in medical references by Brouwers et al. [1] and Armstrong and Gronseth [2]. We identified similarities and differences and characterized what is applicable to visualization design guidelines. This led us to an updated list of 41 properties (organized in the 11 themes identified before) as well as insights into how to phrase and organize the properties. To support the presentation, we further organized the 11 themes into 5 groups.

Derive Packaging approach: We then worked on how the guidelines could be packaged, i.e, represented in ways to make them most useful to designers. We worked on a template that provides a number of questions that a guideline developer should answer for their audience. We then cross-checked this template against the properties. Based on that, we iterated and updated the template to improve coverage and reduce redundancy.

Characterizing next steps: With the above steps, we built the foundation for properties and packaging (template). In the next steps, we want to survey a representative sample of existing guidelines (making use of existing repositories of design guidelines [3, 4] as a starting point) to formatively and summatively evaluate the properties and the template for representing guidelines. We have only begun to discuss the third pillar – the ramifications of how to include our approach into existing design processes.

Our initial results are:

- 11 property themes, summarizing the overall set of xx properties
- A packaging template with 13 prompts
- A set of insights into guideline effectiveness for visualization design

Acknowledgments: Other members of the group who have contributed to the discussions were Alexander Bock and Fabian Beck, as well as the larger group of Dagstuhl participants. Thanks a lot!

References

1 Brouwers, M.C., Kho, M.E., Browman, G.P., Burgers, J.S., Cluzeau, F., Feder, G., Fervers, B., Graham, I.D., Grimshaw, J., Hanna, S.E. and Littlejohns, P., 2010. AGREE II: advancing guideline development, reporting and evaluation in health care. Cmaj, 182(18), pp.E839-E842.

2 Armstrong, M.J. and Gronseth, G.S., 2018. Approach to assessing and using clinical practice guidelines. Neurology: Clinical Practice, 8(1), pp.58-61.

3 Choi, J., Oh, C., Suh, B. and Kim, N.W., 2021, May. Toward a Unified Framework for Visualization Design Guidelines. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-7).

4 Diehl, A., Abdul-Rahman, A., El-Assady, M., Bach, B., Keim, D.A. and Chen, M., 2018. Visguides: A forum for discussing visualization guidelines. EuroVis (Short Papers), 6(7), pp.61-65.

4.3 From Cognition to Context: A Conversation about Technical Approaches, Social Values, and Tradeoffs in Visualization

Miriah Meyer (Linköping University, SE), Lane T Harrison (Worcester Polytechnic Institute, US), Alex Kale (University of Chicago, US), Carolina Nobre (University of Toronto, CA), and Arvind Satyanarayan (MIT – Cambridge, US)

Our working group produced a panel proposal for the IEEE VIS conference as part of our small group efforts at the seminar. This panel was well-received at the conference – with standing room only – and sparked many conversations throughout the week of the conference. As part of our artifacts we created a zine that we distributed at the panel.

4.4 From Paper to Prompt: Teaching AI to Apply the Rules Using AI to extract, adapt, and apply visualization guidelines

Vidya Setlur (Tableau Research – Palo Alto, US), Michael Aupetit (HBKU – Doha, QA), Fabian Beck (Universität Bamberg, DE), Angelos Chatzimparmpas (Utrecht University, NL), Sungahn Ko (POSTECH – Pohang, KR), Kuno Kurzhals (Universität Stuttgart, DE), and Tatiana von Landesberger (Universität Köln, DE)

Despite the abundance of visualization and visual analytics guidelines, users, particularly non-experts, struggle to access, interpret, and apply them effectively. This is due to their fragmented nature, complexity, and nuanced contextual applicability. AI presents a promising proposition by offering the ability to dynamically extract, unify, and adapt these diverse guidelines to specific user needs, tasks, and domains. Key challenges include dealing with the scattered and inconsistent nature of guidelines, ensuring contextual relevance and adaptability, verifying the accuracy and reliability of AI-generated recommendations, and supporting diverse user interaction needs and degrees of expected automation. This opens opportunities for developing AI techniques and tools that automatically extract and structure guidelines, apply them based on user goals and data context, and ensure continual learning and personalization. Research directions include context-aware recommendation systems, multi-agent architectures for modular reasoning, and robust evaluation frameworks to ensure trustworthy and actionable AI-driven visualization support.

We created a zine to summarize our working group ideas.

Participants

- Bon Adriel Aseniero
AUTODESK – Toronto, CA
- Michael Aupetit
HBKU – Doha, QA
- Cindy Xiong Bearfield
Georgia Institute of Technology – Atlanta, US
- Fabian Beck
Universität Bamberg, DE
- Alexander Bock
Linköping University, SE
- Angelos Chatzimparmpas
Utrecht University, NL
- Michael Gleicher
University of Wisconsin-Madison, US
- Lane T Harrison
Worcester Polytechnic Institute, US
- Petra Isenberg
INRIA Saclay – Orsay, FR
- Alex Kale
University of Chicago, US
- Sungahn Ko
POSTECH – Pohang, KR
- Kuno Kurzhals
Universität Stuttgart, DE
- Miriah Meyer
Linköping University, SE
- Carolina Nobre
University of Toronto, CA
- Ghulam Jilani Quadri
University of Oklahoma – Norman, US
- Paul Rosen
University of Utah – Salt Lake City, US
- Arvind Satyanarayan
MIT – Cambridge, US
- Karen Schloss
University of Wisconsin – Madison, US
- Michael Sedlmair
Universität Stuttgart, DE
- Vidya Setlur
Tableau Research – Palo Alto, US
- Cagatay Turkay
University of Warwick – Coventry, GB
- Tatiana von Landesberger
Universität Köln, DE
- Daniel Weiskopf
Universität Stuttgart, DE

Report from Dagstuhl Seminar 25241

Utilising and Scaling the WebAssembly Semantics

Amal Ahmed[*1], Andreas Rossberg[*2], Deian Stefan[*3], Conrad Watt[*4], and Michelle Thalakottur[†5]

1 **Northeastern University – Boston, US.** amal@ccs.neu.edu
2 **München, DE.** rossberg@mpi-sws.org
3 **University of California – San Diego, US.** deian@cs.ucsd.edu
4 **Nanyang TU – Singapore, SG.** conrad.watt@cl.cam.ac.uk
5 **Northeastern University – Boston, US.** michelledaviest@gmail.com

Abstract

WebAssembly (Wasm) is a safe and portable, low level bytecode format used in browsers, IoT applications, cloud, edge, embedded systems and blockchains. Its popularity as a technology for both practically building and theoretically investigating verified and secure systems has been growing rapidly. This Dagstuhl Seminar brought together leading academics and industry representatives involved in Wasm, both as designers, implementers or clients of the technology, to exchange ideas around topics such as tools for formal specification, verified compilation, software fault isolation and language interoperability.

Seminar June 9–13, 2025 – https://www.dagstuhl.de/25241
2012 ACM Subject Classification Software and its engineering → Semantics; Software and its engineering → Runtime environments; Theory of computation → Program semantics
Keywords and phrases Compilers, Formal Methods, JavaScript, Proof Assistants, Runtimes, Software Verification, Webassembly
Digital Object Identifier 10.4230/DagRep.15.6.51

1 Executive Summary

Amal Ahmed (Northeastern University – Boston, US)
Andreas Rossberg (München, DE)
Deian Stefan (University of California – San Diego, US)
Conrad Watt (Nanyang TU – Singapore, SG)

WebAssembly (Wasm) is a safe and portable code format used in a broad variety of computational environments, such as Web browsers, cloud, edge, IoT, embedded systems, and blockchains. As a low-level programming language its instruction set is close to that of physical hardware, yet its semantics enforces memory safety, isolation, and the absence of undefined behavior. A distinguishing feature of Wasm is that its official specification contains a complete formal semantics, based directly on techniques developed and established by the scientific community, and proved sound with machine-verified proofs. Its popularity as a technology for both practically building and theoretically investigating verified and secure systems has hence been growing rapidly.

This Dagstuhl Seminar, brought together all sides interested in Wasm, its formal semantics, and its application to verification and generation techniques. By including academics and

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

industry representatives involved in Wasm, both as designers, implementers or clients of the technology (e.g., compiler writers), we hoped to initiate discussion and new research about evolving the specification, as well as utilizing it for implementing verified systems, programming languages, and new forms of tooling. The main focus was around the following four topics:

Tools for formal specification. Maintaining the current level of rigor in Wasm's specification is an ongoing challenge. While many new features have been proposed for Wasm, the risk is that either the formal specification gets in the way of evolving the language, or that the formalization falls behind.

Verified compilation. We believe that Wasm's formal semantics is an excellent starting point for performing verification in depth for systems built around the language. This applies both to compilation from source languages to Wasm, and compilation from Wasm to native code as it occurs in Wasm engines and their just-in-time compilers.

Software fault isolation. Software fault isolation can be implemented at various granularities and using different techniques, ranging from inline software checks to hardware-based isolation. Such techniques can inform both Wasm's design and its implementations. Wasm can also serve as a vehicle for lightweight isolation, from sandboxing libraries to more ambitious projects, such as a hypervisor or an embedded execution environment.

Language interoperability. In Wasm multi-language interoperation can occur in two ways: first, as the interaction between a language compiled to Wasm and the language in which the host environment operates, and second between multiple different languages compiled to Wasm.

2 Table of Contents

Working groups

3 Overview of Talks

3.1 Cross-language multi-core symbolic execution with Owi and more

Léo Andrès (OCamlPro – Paris, FR)

Joint work of Léo Andrès, Filipe Marques, Pierre Chambart, Arthur Carcano
Main reference Léo Andrès, Filipe Marques, Arthur Carcano, Pierre Chambart, José Fragoso Santos, Jean-Christophe Filliâtre: "Owi: Performant Parallel Symbolic Execution Made Easy, an Application to WebAssembly", Art Sci. Eng. Program., Vol. 9(1), 2024.
URL https://doi.org//10.22152/PROGRAMMING-JOURNAL.ORG/2025/9/3

I'll present new things happening in Owi, a Wasm toolkit featuring a symbolic execution engine. I'll show how we use it to perform automated bug-finding on C, C++, Rust and Zig programs and a bug we found in the Rust stdlib. I'll also present more recent work : how we run Wasm in a Unikernel using MirageOS, the work on owi iso to check Binaryen's optimizations, proof of programs by reusing ACSL (a specification language for C), Weasel (a draft Wasm specification language), support for advanced test coverage criteria (e.g. MCDC), and future work (handling Haskell, TinyGo, OCaml and Guile), build system integration and support for WASI/Component model/Common ABI.

3.2 A Quick Tour on CompCert

Sandrine Blazy (University of Rennes, FR)

URL https://compcert.org/

The CompCert verified C compiler is an open infrastructure for research. It accomplishes a series of 18 passes through 9 intermediate languages. I will explain how their semantics is mechanized and how this facilitates the reasoning required to prove compiler correctness.

3.3 Motoko: Enhanced Orthogonal Persistence

Claudio Russo (DFINITY Foundation – Zürich, CH)

Joint work of Luc Bläser, Claudio Russo, Gabor Greif, Ryan Vandersmith, Jason Ibrahim
Main reference Luc Bläser, Claudio Russo, Gabor Greif, Ryan Vandersmith, Jason Ibrahim: "Smarter Contract Upgrades with Orthogonal Persistence", in Proc. of the 16th ACM SIGPLAN International Workshop on Virtual Machines and Intermediate Languages, VMIL 2024, Pasadena, CA, USA, 20 October 2024, pp. 32–42, ACM, 2024.
URL https://doi.org//10.1145/3689490.3690401

Motoko is a strongly-typed, actor based language with impure functional features that targets the Internet Computer blockchain. I'll give an overview on Motoko's unique support for data persistence across requests and code upgrades (the enhanced bit).

3.4 Automating Coq Mechanization for Webassembly via Spectec

Diego Cupello (Imperial College London, GB) and Philippa Gardner (Imperial College London, GB)

Spectec is a new toolchain for inputting specifications, namely for Wasm. It delivers specification requirements that all come from a "single source of truth", being a new DSL that is designed with human readability and expressivity in mind. I will give a small overview of the prototype pass made to generate Rocq definitions, and give some advances in designing the future structure that allows for generation of Inductive definitions in many interactive theorem provers, including Rcoq, Isabelle, Lean and Agda. Furthermore, I will also give a small proof of concept extension that allows lemmas and theorems to be in part of the DSL, with a mechanism that allows the generation of boilerplate lemmas.

3.5 RichWasm Realizability: Facilitating Formal Specification of ABIs for WebAssembly

Ryan Doenges (Northeastern University – Boston, US)

Joint work of Brianna Marshall, Ryan Doenges, Maxime Legoupil, Lars Birkedal, Amal Ahmed

For programs in different high-level languages to interoperate after translation to WebAssembly, they need a shared Application Binary Interface (ABI) that standardizes data layout, memory management policies, and calling conventions. This means each language has to have an independent definition of how its types relate to the ABI, which leads to incompatibilities and duplication of effort.

To unify these ABI specifications, we describe RichWasm, a low-level IR with high-level types for interoperation between GC'd and manually managed languages, and equip it with a realizability model in Iris. The model interprets RichWasm types as separation logic specifications for WebAssembly terms, while the type system of RichWasm itself is expressive enough to support type-preserving compilation from source languages like ML or L3, a calculus of linear references. Our model is implemented in the Iris-Wasm program logic. To show the model is useful, we prove it is adequate for a memory isolation property. To show the model is not vacuous, we implement a compiler from RichWasm to WebAssembly and prove that it sends terms of type T to inhabitants of the realizability model at T.

This talk describes work in progress.

3.6 WasmGC languages: where are we?

Sébastien Doeraene (EPFL Lausanne, CH)

Joint work of Sébastien Doeraene, Rikito Taniguchi, Tobias Schlatter

A look at the state of WasmGC, in terms of the capabilities it gives to languages ... and what it doesn't give. We will particularly look at questions of interoperability (with JS and with the Component Model) and performance ("unnecessary" casts, late binding).

Interoperability with JavaScript has made leaps with the introduction of WasmGC, as we can now have cycles between GCed structures across JavaScript and Wasm. There are still some gaps, but proposals are on their way to address them. Using the Component Model from a GC language remains a big question mark, especially with the requirements for ubiquitous copies, and unknowns regarding drop semantics.

Performance is fine, provided a decent ahead-of-time optimizer. In some cases, our best efforts are still slower than equivalent JavaScript code. The main performance problems derive from two sources: calls to small JavaScript bridge functions, and virtual/interface method dispatch. We also discuss some smaller, possibly low-hanging fruit.

3.7 Verifying Wasm-to-Native Compilation with Authoritative ISA Specifications

Chris Fallin (F5 – San Jose, US)

Joint work of Chris Fallin, Alexa VanHattum, Michael McLoughlin, Adrian Sampson, Brian Parno, Fraser Brown, Monica Parneshi
Main reference Alexa VanHattum, Monica Pardeshi, Chris Fallin, Adrian Sampson, Fraser Brown: "Lightweight, Modular Verification for WebAssembly-to-Native Instruction Selection", in Proc. of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024, pp. 231–248, ACM, 2024.
URL https://doi.org//10.1145/3617232.3624862
Main reference Michael McLoughlin, Ashley Sheng, Chris Fallin, Bryan Parno, Fraser Brown, Alexa VanHattum: "Scaling Instruction-Selection Verification against Authoritative ISA Semantics", Proc. ACM Program. Lang., Vol. 9(OOPSLA2), Association for Computing Machinery, 2025.
URL https://doi.org//10.1145/3764383

When compiled to native execution, WebAssembly's guarantees are only as strong as the compiler's. Subtle wrong-code bugs, including in instruction selection, have introduced serious security vulnerabilities. In this talk, we'll describe our system for lightweight, modular verification of instruction-lowering rules in Wasmtime's Cranelift compiler backend. We now automatically derive machine-specific specifications from vendor-provided ISA semantics (via partial evaluation of ASL) for ARM aarch64. We'll discuss the particular challenges of handling partially-symbolic instructions, e.g., for immediate operands. Our design also benefits from lightweight state modeling and compositional reasoning over chains of rewrite rules. We reproduce known bugs, including one of the most severe security bugs in Wasmtime's history (from 2023, concurrent with the last Wasm Dagstuhl).

3.8 Binaryen IR and Type System

Thomas Lively (Google – Mountain View, US)

Joint work of Thomas Lively, Alon Zakai, The WebAssembly Community Group
URL https://github.com/webassembly/binaryen

Binaryen is a widely used WebAssembly optimizer, but its IR design predates the decision to make WebAssembly a stack machine. This, among other considerations, leads to some interesting differences between the standard WebAssembly type system and the one used in Binaryen. This talk will be a brief tour of those differences and their consequences.

3.9 A Logical Interface for Garbage Collectors Implemented in WebAssembly

Brianna Marshall (Northeastern University – Boston, US)

Joint work of Brianna Marshall, Ryan Doenges, Maxime Legoupil, Owen Duckham, Ari Prakash, Lars Birkedal, Amal Ahmed

As part of defining an ABI for RichWasm, we must specify how garbage collected references are encoded in WebAssembly. Because RichWasm requires that the garbage collector can free linear references owned by a collected object, the target GC needs to support finalizers, which rules out the Wasm GC extension. We describe an approach for "rolling our own GC" in Iris-Wasm that encapsulates the physical memory owned by the GC behind a separation logic interface. The mutator instead owns blocks of abstract memory, which can be converted temporarily to the underlying physical resources so that the memory can be used directly between invocations of the GC.

3.10 Execution-Aware Program Reduction for WebAssembly via Record and Replay

Michael Pradel (Universität Stuttgart, DE)

Joint work of Michael Pradel, Doehyun Baek, Daniel Lehmann, Ben L. Titzer, Sukyoung Ryu
Main reference Doehyun Baek, Daniel Lehmann, Ben L. Titzer, Sukyoung Ryu, Michael Pradel: "Execution-Aware Program Reduction for WebAssembly via Record and Replay", CoRR, Vol. abs/2506.07834, 2025.
URL https://doi.org//10.48550/ARXIV.2506.07834

WebAssembly (Wasm) programs may trigger bugs in their engine implementations. To aid debugging, program reduction techniques try to produce a smaller variant of the input program that still triggers the bug. However, existing execution-unaware program reduction techniques struggle with large and complex Wasm programs, because they rely on static information and apply syntactic transformations, while ignoring the valuable information offered by the input program's execution behavior. We present RR-Reduce and Hybrid-Reduce, novel execution-aware program reduction techniques that leverage execution behaviors via record and replay. RR-Reduce identifies a bug-triggering function as the target function, isolates

that function from the rest of the program, and generates a reduced program that replays only the interactions between the target function and the rest of the program. Hybrid-Reduce combines a complementary execution-unaware reduction technique with RR-Reduce to further reduce program size. We evaluate RR-Reduce and Hybrid-Reduce on 28 Wasm programs that trigger a diverse set of bugs in three engines. On average, RR-Reduce reduces the programs to 1.20% of their original size in 14.5 minutes, which outperforms the state of the art by 33.15x in terms of reduction time. Hybrid-Reduce reduces the programs to 0.13% of their original size in 3.5 hours, which outperforms the state of the art by 3.42x in terms of reduced program size and 2.26x in terms of reduction time. We envision RR-Reduce as the go-to tool for rapid, on-demand debugging in minutes, and Hybrid-Reduce for scenarios where developers require the smallest possible programs.

3.11 WasmCert-Rocq and the Designs for Future Mechanisations

Xiaojia Rao (Imperial College London, GB)

URL https://github.com/WasmCert/WasmCert-Coq

WasmCert-Rocq has evolved much further since its initial formulation of Wasm 1.0. The mechanisation has shifted its target version of Wasm from 1.0 to 2.0 and beyond, and the scope of work included in the mechanisation has also vastly expanded from the initial soundness results. In this talk, I will give an overview of the current structure of the Rocq mechanisation, which formalises Wasm 2.0 + tail call, implementing the subtyping system from future proposals. I will then highlight some of the most interesting designs devised since the inception of the mechanisation, including the progressful interpreter for proof consolidation, an efficient persistent data structure for extracting memory representations without resolving fully to monadic state operations, and some consideration in extraction for convenient host interoperations. These designs can stay relevant in producing an efficient verified reference interpreter in the future, when a SpecTec-generated mechanisation comes to fruition. The extracted runtime now passes the Wasm 2.0 core test suite fully.

3.12 A Quick Tutorial on SpecTec

Andreas Rossberg (München, DE)

Main reference Dongjun Youn, Wonho Shin, Jaehyun Lee, Sukyoung Ryu, Joachim Breitner, Philippa Gardner, Sam Lindley, Matija Pretnar, Xiaojia Rao, Conrad Watt, Andreas Rossberg: "Bringing the WebAssembly Standard up to Speed with SpecTec", Proc. ACM Program. Lang., Vol. 8(PLDI), pp. 1559–1584, 2024.
URL https://doi.org//10.1145/3656440

SpecTec is the new DSL and tool chain for authoring the Wasm spec. It can be used to generate various output formats from a formal definition, like Latex rules, English prose, Coq definitions, etc. This presentation gave a brief overview of the SpecTec language itself, using a small fragment of Wasm as a running example and showing how a specification can be created with it.

3.13 Propagating Host Types for Multi-Language Static Analysis of WebAssembly

Michelle Thalakottur (Northeastern University – Boston, US)

Joint work of Thalakottur, Michelle; Garg, Harshit; Mane Siddhant; Fallin Chris; Ahmed Amal; Tip, Frank

Current state of the art static analysis tools analyze WebAssembly binaries in isolation, that is, they perform a closed-world analysis where they make no assumptions about the Javascript client interacting with the WebAssembly binary, resulting in significant loss of precision. Inspired by real-world invocation patterns of WebAssembly by compiler-generated Javascript wrappers, we design a refinement type system over WebAssembly. We infer refined WebAssembly types by generating and solving constraints and use them to generate call graphs that are more precise than those generated from a close-world analysis.

3.14 Engine Interfaces for Wasm Instrumentation

Ben L. Titzer (Carnegie Mellon University – Pittsburgh, US)

While Wasm enjoys excellent performance on many platforms due to investment in engine optimization, challenges remain to improve Wasm as a development target. In particular, the lack of standardized debugging and profiling tools across engines has led to fragmentation and less than ideal developer experience. Wasm lags behind other environments like the JVM by a wide margin. In this talk I will outline steps that we've taken to define an engine interface for programmable instrumentation with low overhead and (hopefully) low engine implementation effort.

3.15 An Adjoint Separation Logic for the Wasm Call Stack

Andrew Wagner (Northeastern University – Boston, US)

Joint work of Andrew Wagner, Zachary Eisbach, Amal Ahmed

We propose a pair of adjoint modalities for separation logics over stacks of resources. This abstraction guarantees the encapsulation expected of a stack discipline in a completely local, small footprint style. In the context of Wasm, these modalities can alleviate the burden of explicit threading around a monolithic resource for each activation frame.

3.16 Future of concurrency in Wasm

Conrad Watt (Nanyang TU – Singapore, SG)

Main reference Conrad Watt, Concurrency in WebAssembly, ACM Queue Volume 23 Issue 3
URL https://dl.acm.org/doi/abs/10.1145/3747201.3746173

WebAssembly has basic support for concurrency through the use of shared memories and the corresponding SharedArrayBuffer in JavaScript. However, because these concurrency capabilities only support the sharing of data in a bytewise representation, some language with richer concurrency features run into expressivity limitations when compiling to WebAssembly. This talk discusses these limitations, and sketches ongoing work within the WebAssembly standards community to extend the language with more powerful concurrency capabilities. Of particular interest are "shared functions", which would give WebAssembly the capability to share executable code cross-thread.

3.17 Towards Performant Static Analysis of WebAssembly via Staging and Continuations

Guannan Wei (INRIA – Paris, FR & Tufts University – Medford, US)

Joint work of Guannan Wei, Dinghong Zhong, Alex Bai

The official WebAssembly specification provides a small-step reduction semantics. However, this semantics is complicated by additional "administrative instructions" to account for WebAssembly's structured control-flow constructs. As a result, the semantics is not compositional and is not amenable for partial evaluation or staging.

In this talk, I presented an alternative natural semantics for WebAssembly in continuation-passing style (CPS), which can be implemented as a concise, compositional, and tail-recursive definitional interpreter. By using continuations from the meta-language, this approach streamlines the semantics and eliminates the need for administrative instructions.

I also discussed our ongoing work using this CPS semantics as a foundation for efficient concolic execution engines of WebAssembly, building on the first Futamura projection. Finally, I outlined future directions, including the use of SpecTec to specify interderivable semantics by program transformation.

3.18 WEST: Generating Wasm Test Cases based on SpecTec

Dongjun Youn (KAIST – Daejeon, KR)

Joint work of Dongjun Youn, Sukyoung Ryu, Wonho Shin

We present WEST (WebAssembly Specification-based Testing), a framework that automatically produces Wasm test cases from mechanized specifications written in SpecTec. Given any full or subset variant of the Wasm specification as input, WEST aims to systematically generate test programs that conform to the input grammar and validation rules, and capture the runtime behavior defined by its execution semantics.

4 Working groups

4.1 Interoperability

Sébastien Doeraene (EPFL Lausanne, CH)

This working group focused on two broad areas of interest related to interoperability between languages on the Wasm platform. First, the run-time cost of cross-language boundaries. Second, the interaction of various type systems.

We discussed what seems to be the biggest factor of the run-time cost of interoperability: extra copies of data. Those copies are motivated by a combination of a) differences in data layout (including GC versus linear models) and b) protections of each language's abstractions. We discussed the need for a unified semantic framework. It could provide the necessary ground for enforcing abstractions across languages, or at least a common denominator of the most important ones. If we are willing to recompile modules when combining them, a unified framework may also be used to specialize modules to various data layouts.

The interactions between type systems are perhaps even trickier, with no clear path forward. Here as well, some notion of common denominator could be used. RichWasm was discussed at length as a possible foundation for type systems to compile down to. Generics were brought up as a particularly hard topic. IDLs offer a different look at the problem space: rather than compiling language type systems down to a common ground, we can "lift up" IDL definitions into source languages.

As takeaway from the working group, it is clear that interoperability is a hard topic. Rather than finding an all-encompassing solution, the most promising path forward seems to take concrete pairs of languages, and find dedicated solutions for those. With time, we may be able to generalize the learning outcomes of several approaches. The "common denominator" was a recurring concept. There is some inherent tension between the various levels at which it could be placed. WasmGC might provide a better foundation for an interoperability layer than the traditional linear-memory-based approaches. Regardless of the approach taken, over time we will need a set of building blocks that languages can leverage.

4.2 Garbage Collection and Dynamic Languages

Chris Fallin (F5 – San Jose, US)

A working group discussed two topics related to high-level language compilation to WebAssembly: the Wasm GC proposal (modes of use, potential extensions) and the handling of dynamism (dynamic types and other kinds of highly-dynamic language semantics).

We first discussed garbage collection adoption in several Wasm guest language compilers, and received an experience report from an author of the ScalaWasm project, including a discussion of the compilation strategy for interfaces, vtables, and approaches that could avoid a sparse layout (slots for every possible interface). We discussed several possible extensions, including those named during the GC proposal development process as post-MVP features: sum types, trailing arrays, intersection types, support for inner pointers and inlined objects

(as supported by languages such as Go), finalizers and weak-references, and others. Consensus was that there are use-cases for all of these features, but some of them may be more difficult than others to support in existing Wasm engines. In particular, we discussed the impact that Web engines' existing JS garbage collector semantics may have: for example, Wasm GC finalizers may be constrained to JS semantics (postmortem finalizers, rather than premortem finalizers that have an opportunity to re-create an edge to an object to retain it), and interior pointers may be difficult to support. We discussed whether formal verification techniques, such as MS-Wasm, may be useful to achieve safer garbage collection implementation; and several sandboxing techniques were also discussed (including engines such as Wasmtime that build GC storage on top of an untrusted Wasm linear memory internally). We discussed test coverage in the official spec test suite and agreed that it should be improved.

Next, we transitioned to a discussion of dynamic language implementation, focusing on inline caches (ICs) as a means to optimize behavior that cannot be known until runtime. We held an extensive discussion of the approach that an existing JavaScript-to-Wasm compilation (using SpiderMonkey and partial evaluation in the StarlingMonkey engine) takes to implement ICs, using an ahead-of-time-collected corpus of fastpath ICs and function pointers to these ICs that are updated. We discussed whether it might make sense to include a feature in core Wasm to allow for small functions with fast hot-patching, and perhaps first-class visibility of IC chains. We concluded that work should first focus on making use of newer Wasm features to reduce the function-call overhead seen in existing ICs (e.g., typed non-nullable function references), combined with targeted engine optimizations (e.g., minimizing the fixed overhead of function prologues and/or pushing function frame creation to cold-paths, similar to how native ICs work) and this should in theory be sufficient to match native code generated by typical dynamic language JITs.

4.3 Performance and Benchmarking

Daniel Lehmann (Google – München, DE)

A small working group discussed topics related to WebAssembly benchmarking and performance, touching on measurement methodology, relevant metrics, tools employed by developers, and different audiences for performance work.

The first observation is that benchmarking and performance measurement methodology varies widely. One group, e.g., the Scala.js developer team and likely other Wasm toolchains focus for now on exercising different language features via microbenchmarks or synthetic workloads. One point of interest for that group is JavaScript interop performance (e.g., passing strings or the language's number type across the Wasm-JS boundary). They don't have large workloads yet that would mimic a full, realistic application. The Ocaml toolchain by Tarides however is in fact evaluated on micro and macro benchmarks, e.g., provided by Jane Street. Since those are closed source or proprietary, it's hard to share them across the Wasm community. Finally, other teams such as the Google V8 team do systematic benchmarking on their CI/testing infrastructure, with automated performance alerts on regressions, larger workloads, and comparisons across different Wasm implementations. Toolchain and engine developers alike are searching for realistic, larger workloads to evaluate performance on, but they are hard to come by, especially during early stages of a toolchain or feature.

A second theme were differences as to what constitutes good performance. For some web applications, it might be peak performance or startup latency or memory usage. In particular the embedded use case of Wasm at Siemens is different in that it focuses on tight instruction budgets, because Wasm is used in control loops with a fixed time window. For those applications inconsistent performance is worse than simply slow execution speeds.

A third question is who is the audience for performance work and benchmarking, or put differently: where to optimize if performance goals are not met. That can be application developers working on the source language code, e.g., C++ or Scala compiled to Wasm; it can be toolchain authors, such as the compilers from Scala to Wasm or Emscripten or Binaryen; and finally it can be engine developers with optimizing JIT compilers.

An open idea during the working group was domain specific acceleration. E.g., similar to the JS-string-builtins proposal, should there be other intrinsics or fast imported functions for AES-NI/crypto primitives? One open question for the Wasm community could be to come up with a generic "builtin" mechanism, which may be polyfilled by a slow, software/scalar implementation as a Wasm module (similar to how JS-string-builtins are polyfillable in JavaScript host code).

Finally, there was a concrete call to action for the Wasm community, namely to engage more among/with the community (consisting of application, toolchain, and engine developers) to collect best practices and interesting workloads for benchmarking, since there is no centralized place for such information as of today. A WebAssembly benchmarking community subgroup already exists but is dormant, it could be revived for that purpose. One toolchain developer also wished for better documentation on what Wasm code is well optimizable by engines and which performance properties one could expect, similar to documents of native micro-architectures.

4.4 SpecTec and Mechanisation

Andreas Rossberg

Main reference Dongjun Youn, Wonho Shin, Jaehyun Lee, Sukyoung Ryu, Joachim Breitner, Philippa Gardner, Sam Lindley, Matija Pretnar, Xiaojia Rao, Conrad Watt, Andreas Rossberg: "Bringing the WebAssembly Standard up to Speed with SpecTec", Proc. ACM Program. Lang., Vol. 8(PLDI), pp. 1559–1584, 2024.
URL https://doi.org//10.1145/3656440

SpecTec is the new specification language for the official WebAssembly definition that grew out of the previous Dagstuhl on WebAssembly. This session was concerned with possible new use cases, open problems, as well as possible improvement. We discussed of various ideas, from near-term to very ambitious.

Various suggestions were made for other artefacts that could potentially be generated from SpecTec, such as different flavours of interpreters (executing the reduction semantics, parsers, symbolic interpreters, CPS transforms), more tests, or other IRs (e.g., for tools like Binaryen). A prover backend might also be able to generate tests for the semantic specification itself, as an aid to language and proposal designers before all proofs are pushed through by mechanisation experts. Such a generator could take advantage of the stated meta-theoretical properties.

A problem for interpreters is the handling of non-determinism in the specification. While it may be desirable to explore the entire space of possible results (e.g., with abstract interpretation), such an attempt is likely to be computationally infeasible in the presence

of features like the relaxed memory model. A simple fallback would be random execution (with a reproducible seed), although that cannot capture relaxed executions that involve reorderings of memory accesses.

Other topics of interest were the integration with spec update process, especially when certain backends lag behind, customising and subsetting specifications, and extending SpecTec to include theorem statements, in a way that would be abstract enough for different mechanisation backends. SpecTec's general expressiveness was briefly discussed (e.g., the availability of evaluation contexts, binders, quantifiers, or judgements as expressions), as well as its IL semantics and possible meta-theory, a more robust and configurable Il-to-AL translation, and paths towards the ideal backend modularisation.

4.5 Static Analysis

Michelle Thalakottur (Northeastern University – Boston, US), Léo Andrès (OCamlPro – Paris, FR), Alexander Bai (MPI für Software Systems – Saarbrücken, DE), Thomas Lively (Google – Mountain View, US), Marco Patrignani (University of Trento, IT), and Guannan Wei (INRIA – Paris, FR & Tufts University – Medford, US)

We discussed what each of us were interested in, which was, performant static analyses, composable static analyses and multi-language analyses. We talked about handling various proposals and how non-determinism is a hard open problem. We also talked about how modeling system calls and calls to JavaScript is a problem faced by every static analysis tool. We all agreed that compiler-produced annotations or hints that are preserved from the source and embedded in custom sections in the Wasm binary, would be useful for a more precise analysis. Thomas Lively gave us an overview of how binaryen, a WebAssembly toolchain from Google, handles the Javascript environment in its static analysis, the different analyses being done in binaryen and what they would like to see implemented or improved. We talked about the difficulties we face while productionizing research static analysis tools. We ended by talking about the possibility of using an analysis platform like CodeQL for multi-language analysis.

4.6 Instrumentation

Ben L. Titzer (Carnegie Mellon University – Pittsburgh, US), Chris Fallin (F5 – San Jose, US), Ralph Squillace (Microsoft – Redmond, US), and Thomas Trenner (Siemens AG – Nürnberg, DE)

This working group focused on three industrial-relevant topics: debugging, introspection, and replay.

Key issues raised were the developer experience, such as IDE integration, and how difficult it can be to debug containers. Since many scenarios deploy Wasm code inside a a container, this greatly improves the challenge. For dynamic languages whose runtimes

already support remote debugging over a socket, the experience seems to somewhat work. But there is diversity here; every language has its own tooling, workflow, APIs, etc for doing this. Microsoft has had some success getting debugging to work for StarlingMonkey (JavaScript) and Python, but other languages don't have this yet.

The working group discussed debugging different semantic levels: the machine (ISA) level, the Wasm bytecode level, and the source level. Different engines have different support. E.g. wasmtime can allow debugging at the machine level in some situations, bytecode at others, and limited support for WasmDWARF.

Other issues that were raised were the inability to debug live cyberphysical systems (e.g. cannot step a $100k machine). There are IP issues about debugging code on some platforms.

The group discussed remote debugging protocols and replay. More standardization and standard interfaces within the Wasm community would help. Members considered syncing again on how to design remote interfaces, e.g. to dig into how the debugging protocols work in VSCode.

4.7 Concurrency in Wasm

Conrad Watt (Nanyang TU – Singapore, SG)

This working group was convened to discuss ways to support the compilation of concurrent code to Wasm, and the current standards body efforts under the "shared-everything threads" proposal to extend the concurrency capabilities of Wasm. The group discussed several topics in detail.

(1) some source languages have the capability to share dynamically-loaded executable code between threads (e.g. C through the dlopen system call). Wasm's support for this capability is currently poor and an extension to support Wasm-level shared functions would alleviate this issue.

(2) some embedded system implementations of Wasm may stretch the limitations of the standard and in particular the standard's specification of host environment interaction in order to provide more powerful concurrency capabilities. After some discussion, it was tentatively determined that many of these behaviours could actually be interpreted as standards-compliant, providing reassurance that the ecosystem can be kept somewhat unified.

(3) as Wasm transitions to a new specification infrastructure based on the SpecTec tool and DSL, new threads features being added to Wasm will require SpecTec to be extended. This represents an opportunity to ensure threads are specified precisely, as well as a hazard that SpecTec extension for this feature may prove to be onerous.

4.8 Testing

Dongjun Youn (KAIST – Daejeon, KR)

In the Wasm testing breakout session, various aspects of the topic were discussed. Both academia and industry are actively working on improving Wasm testing through multiple approaches. Current testing methods include random testing, where random inputs are generated for testing; mutation testing, which involves intentionally altering Wasm code to test the system's response; and using SpecTec as oracle for test generation, or using differential testing. There is also an ongoing effort to reduce redundancy and improve test performance through dynamic checks. New ideas and approaches presented during the meeting included generating negative tests, which focus on creating invalid or incorrect syntax, and testing for invariants to ensure certain conditions hold throughout execution. Property-based testing leveraging SpecTec for verification, was also discussed as a promising method. The potential use of large language models (LLMs) for generating testing was also explored. Evaluation of testing approaches emphasized the importance of test coverage. A suggestion was made to require a certain level of coverage of specification or reference interpreter before tests are officially adopted. Additionally, mutation testing was proposed, where faults are deliberately injected into the specification to test if existing tests can identify these errors. Several open problems and limitations were identified, including testing interoperation with other components like JavaScript, WASI, and polyfills. The issue of test format incompatibility was highlighted, as well as the performance tradeoffs in dynamic checks and challenges posed by nondeterminism in the tests. In conclusion, the discussion emphasized the importance of continuous cooperation between academia and industry to address these challenges and advance Wasm testing.

Participants

- Amal Ahmed
Northeastern University – Boston, US
- Léo Andrès
OCamlPro – Paris, FR
- Alexander Bai
MPI für Software Systems – Saarbrücken, DE
- Sandrine Blazy
University of Rennes, FR
- Zilin Chen
Nanyang TU – Singapore, SG
- Diego Cupello
Imperial College London, GB
- Ryan Doenges
Northeastern University – Boston, US
- Sébastien Doeraene
EPFL Lausanne, CH
- Chris Fallin
F5 – San Jose, US
- Philippa Gardner
Imperial College London, GB
- Maxime Legoupil
Aarhus University, DK
- Daniel Lehmann
Google – München, DE
- Sam Lindley
University of Edinburgh, GB
- Thomas Lively
Google – Mountain View, US
- Brianna Marshall
Northeastern University – Boston, US
- Tyler McMullen
Fastly – San Francisco, US
- Lucy Menon
London, GB
- Marco Patrignani
University of Trento, IT
- Jean Pichon-Pharabod
Aarhus University, DK
- Michael Pradel
Universität Stuttgart, DE
- Xiaojia Rao
Imperial College London, GB
- Andreas Rossberg
München, DE
- Claudio Russo
DFINITY Foundation – Zürich, CH
- Ralph Squillace
Microsoft – Redmond, US
- Deian Stefan
University of California – San Diego, US
- Michelle Thalakottur
Northeastern University – Boston, US
- Ben L. Titzer
Carnegie Mellon University – Pittsburgh, US
- Thomas Trenner
Siemens AG – Nürnberg, DE
- Marco Vassena
Utrecht University, NL
- Jérôme Vouillon
Tarides – Paris, FR
- Andrew Wagner
Northeastern University – Boston, US
- Conrad Watt
Nanyang TU – Singapore, SG
- Guannan Wei
INRIA – Paris, FR & Tufts University – Medford, US
- Chris Woods
Siemens – Princeton, US
- Dongjun Youn
KAIST – Daejeon, KR

Report from Dagstuhl Seminar 25242

Testing Program Analyzers and Verifiers

Maria Christakis[*1], Alastair F. Donaldson[*2], John Regehr[*3], and Thodoris Sotiropoulos[*4]

1 **TU Wien, AT.** maria.christakis@tuwien.ac.at
2 **Imperial College London, GB.** alastair.donaldson@imperial.ac.uk
3 **University of Utah – Salt Lake City, US.** regehr@cs.utah.edu
4 **ETH Zürich, CH.** theodoros.sotiropoulos@inf.ethz.ch

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25242 "Testing Program Analyzers and Verifiers". Program analyzers and verifiers are routinely employed during software development to prevent and detect faults. In this seminar, we examine the impact of faults within these tools, distinguishing between those that are critical and those that are less severe. We also explore and discuss state-of-the-art techniques for uncovering faults in program analyzers and verifiers, their connections to related domains such as compiler testing, and potential future directions for improving their reliability.

Seminar June 9–12, 2025 – https://www.dagstuhl.de/25242
2012 ACM Subject Classification Software and its engineering → Software testing and debugging; Software and its engineering → Software verification
Keywords and phrases formal methods, program analysis, static analysis, testing, verification
Digital Object Identifier 10.4230/DagRep.15.6.69

1 Executive Summary

Thodoris Sotiropoulos (ETH Zürich, CH)

In an era where software pervades every facet of modern life, ensuring the correctness of software systems is of paramount importance. Program analyzers and verifiers are integral ingredients of this endeavor. They provide developers with essential tools to identify and prevent potential software faults before they impact production systems and end users. However, just like all software, analyzers and verifiers are not free from faults. Faults in analysis and verification tools can potentially undermine the entire software ecosystem, leading to missed vulnerabilities, wasted development efforts, and more.

The goal of this Dagstuhl Seminar was to (1) identify the challenges associated with the problem of ensuring the reliability of program analyzers and verifiers, (2) discuss potential ways to address these challenges, and (3) connect both practitioners as well as researchers working in this domain. In particular, the seminar aimed to bring together experts in program analysis, verification, automated testing, and formal methods. The discussion focused on three themes:

- **Severity of faults within program analysis and verification tools**: This involved a detailed examination of how different types of faults can impact various user groups (e.g., end users, software developers) by taking into consideration the context and the

[*] Editor / Organizer

Testing Program Analyzers and Verifiers, *Dagstuhl Reports*, Vol. 15, Issue 6, pp. 69–83
Editors: Maria Christakis, Alastair F. Donaldson, John Regehr, and Thodoris Sotiropoulos
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

intended users of the analysis. For example, we elucidated essential properties required in analyses (e.g., soundness for safety-critical software), while discussing which properties may be less critical.

- **Automated generation of test inputs for analyzers and verifiers**: We aimed to thoroughly discuss the challenge of test input generation for finding faults in program analysis and verification, and explored whether existing program generators used in other contexts (e.g., compiler testing) could be potentially applied to validate analyzers/verifiers. For example, program analysis and verification tools are routinely used to detect a huge variety of semantic errors (e.g., buffer overflows, type errors, integer overflows, and many more). A key technical challenge with this is the generation of programs that exhibit interesting semantic errors that are supposedly to be caught by the analyzer/verifier under test.
- **Test oracles to validate program analyzers and verifiers**: Automatically testing an analyzer or verifier requires a test oracle to determine whether it functions as expected. Given that the majority of program analyzers/verifiers lack a specification and are not standardized, this introduces a significant challenge in determining whether their behavior is correct. The goal was to discuss potential test oracles for different types of faults within program analyzers and verifiers.

The aforementioned discussions were realized through a combination of talks and panel discussions. The seminar began with a brief introduction of the attendees, followed by a keynote by John Regehr outlining the seminar's goals. Afterwards, 14 participants from both academia and industry presented work related to the seminar's scope, including novel test oracles for detecting faults in program analyzers and verifiers, testing in emerging domains (e.g., quantum platforms), and new program generation techniques tailored to analyzers.

The seminar schedule was intentionally open and flexible, encouraging attendees to propose discussion topics. Two dedicated discussion panels were held. The first panel focused on whether proactive testing of program analyzers and verifiers is worthwhile. The debate was structured around opposing viewpoints: John Regehr argued in favor of testing analyzers and verifiers, while Alastair Donaldson presented arguments against it.

The second panel covered a broader range of themes, including the significance of faults in analyzers and verifiers, the role of AI in testing, the problem of program generation saturation, and methods for assessing the performance of analyzers and verifiers.

2 Table of Contents

3 Overview of Talks

3.1 Better Fuzzing via Grammar Mutation and Repair

Cristian Cadar (Imperial College London, GB)

Joint work of Bachir Bendrissou, Alastair Donaldson, Cristian Cadar
Main reference Bachir Bendrissou, Cristian Cadar, Alastair F. Donaldson: "Grammar Mutation for Testing Input Parsers", ACM Trans. Softw. Eng. Methodol., Vol. 34(4), pp. 116:1–116:21, 2025.
URL https://doi.org//10.1145/3708517

In this talk, I presented our recent and ongoing work on enhancing grammar-based fuzzing via grammar mutation and repair.

Our project GMutator aims to generate edge inputs, particularly those incorrectly accepted by a program, via the novel concept of grammar mutation – where the grammar is first mutated before being used for fuzzing.

Our project AFLRepair combines grammar-based fuzzing with greybox fuzzing. To avoid input mutations leading to mostly invalid inputs, we combine standard byte-level mutations with a repair stage.

Both our projects found bugs that are out of reach of prior approaches.

3.2 Testing Equivalence Checkers

Alastair F. Donaldson (Imperial College London, GB)

Joint work of Michalis Pardalos, Alastair F. Donaldson, Emiliano Morini, Laura Pozzi, John Wickerson
Main reference Michalis Pardalos, Alastair F. Donaldson, Emiliano Morini, Laura Pozzi, John Wickerson: "Who checks the checkers? Automatically finding bugs in C-to-RTL formal equivalence checkers", in Proc. of the DVCon Europe 2024; Design and Verification Conference and Exhibition Europe, pp. 39–44, 2024.
URL https://doi.org//10.30420/566438006

C-to-RTL (register-transfer level) formal equivalence checkers (ECs) allow hardware implementations to be compared against software specifications. Thanks to their complete state-space coverage, ECs are trusted to authorise design sign-off. Therefore, ridding ECs of bugs is a top priority. In pursuit of this goal, we have developed Equifuzz, a technique and tool for randomized testing (fuzzing) of SystemC-to-RTL ECs. Equifuzz uses knowledge of SystemC semantics to generate rich designs that are known to be equivalent to trivial RTL designs. It has uncovered 7 unsoundness bugs in major commercial ECs (where the EC claimed equivalence incorrectly), and 5 incompleteness bugs (where the EC failed to prove equivalence between equivalent designs), all of which have been confirmed by the tool vendors. The fact that Equifuzz has been able to find serious bugs in extensively tested, major commercial ECs demonstrates that fuzzing is a valuable complement to the handcrafted tests that EC developers use as standard.

3.3 Search+LLM-based Testing for ARM Simulators

Karine Even-Mendoza (King's College London, GB)

Main reference Karine Even-Mendoza, Héctor D. Menéndez, William B. Langdon, Aidan Dakhama, Justyna Petke, Bobby R. Bruce: "Search+LLM-Based Testing for ARM Simulators", in Proc. of the 47th IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice, SEIP@ICSE 2025, Ottawa, ON, Canada, April 27 – May 3, 2025, pp. 469–480, IEEE, 2025.
URL https://doi.org//10.1109/ICSE-SEIP66354.2025.00047

In order to aid quality assurance of large complex hardware architectures, system simulators have been developed. However, such system simulators do not always accurately mirror what would have happened on a real device. A significant challenge in testing these simulators comes from the complexity of having to model both the simulation and the infinite number of software that could be run on such a device.

Our previous work introduced SearchSYS, a testing framework for software simulators. SearchSYS leverages a large language model for initial seed C code generation which is then compiled, and the resultant binary is fed to a fuzzer. We then use differential testing by running the outputs of fuzzing on real hardware and a system simulator to identify mismatches.

In this talk, we present and discuss our solution to the problem of testing software simulators, using SearchSYS to test the gem5 VLSI digital circuit simulator, employed by ARM to test their systems. In particular, we focus on the simulation of the ARM silicon chip Instruction Set Architecture (ISA). SearchSYS can create test cases that activate bugs by combining LLMs, fuzzing, and differential testing. Using only LLM, SearchSYS identified 74 test cases that activated bugs. By incorporating fuzzing, this number increased by 93 additional bug-activating cases within 24 hours. Through differential testing, we identified 624 bugs with LLM-generated test cases and 126 with fuzzed test inputs. Out of the total number of bug-activating test cases, 4 unique bugs have been reported and acknowledged by developers. Additionally, we provided developers with a test case suite and fuzzing statistics, and open-sourced SearchSYS.

3.4 Constraint-Based Test Oracles for Program Analyzers

Markus Fleischmann (TU Wien, AT), Maria Christakis (TU Wien, AT), Anastasia Isychev (TU Wien, AT), David Kaindlstorfer (TU Wien, AT), and Valentin Wüstholz (Consensys – Wien, AT)

Main reference Markus Fleischmann, David Kaindlstorfer, Anastasia Isychev, Valentin Wüstholz, Maria Christakis: "Constraint-Based Test Oracles for Program Analyzers", in Proc. of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 - November 1, 2024, pp. 344–355, ACM, 2024.
URL https://doi.org//10.1145/3691620.3695035

Program analyzers implement complex algorithms and, as any soft- ware, can contain bugs. Bugs in their implementation may lead to analyzers being imprecise and failing to verify safe programs, i.e., programs with no reachable error locations; or worse, analyzer bugs may lead to reporting unsound results by verifying unsafe programs, i.e., programs with reachable error locations.

In this paper, we propose a method to detect such bugs by gen- erating constraint-based test oracles for analyzers. We re-purpose and extend Fuzzle, a tool for benchmarking fuzzers, in a tool called Minotaur. Minotaur generates C programs from SMT constraints, and based on the satisfiability of the constraints, derives whether the generated programs are safe or unsafe. For instance, for an unsafe program, an analyzer under test contains a soundness issue if it proves it safe. Using Minotaur, we found 30 unique soundness and precision issues in 11 well-known analyzers that reason about reachability properties.

3.5 Checkification: Testing Your (Static Analysis) Truths

Manuel Hermenegildo (IMDEA Software Institute – Pozuelo de Alarcón, ES & UPM – Madrid, ES)

Joint work of Manuel Hermenegildo, Ignacio de Casso, Daniela Ferreiro, Pedro Lopez-Garcia, and Jose F. Morales
Main reference Daniela Ferreiro, Ignacio Casso, José F. Morales, Pedro López-García, Manuel V. Hermenegildo: "Checkification: A Practical Approach for Testing Static Analysis Truths", CoRR, Vol. abs/2501.12093, 2025.
URL https://doi.org//10.48550/ARXIV.2501.12093

In this talk we will present and demo our "checkification" approach: a simple, automatic method for testing static analyzers. Broadly, it consists in checking that the properties inferred statically are satisfied dynamically. The main advantage of checkification lies in its simplicity, specially when framed within the Ciao assertion-based validation framework, which implements a blend of static and dynamic assertion checking.

We will demonstrate how in this setting analysis results can be tested with little effort by combining, via a simple program transformation, the basic components that comprise the framework itself: 1) the static analyzer, which outputs its results as the original program source with assertions interspersed; 2) the assertion run-time checking mechanism, which instruments a program to ensure that no assertion is violated at run time; 3) the random test case generator, which generates random test cases satisfying the properties present in assertion preconditions; and 4) the unit-testing framework, which executes those test cases. We will show the interaction of these components while checking the results of the CiaoPP abstract interpretation-based static analyzer, for several abstract domains for different properties, analysis (fixpoint) algorithms, etc.

3.6 Interrogation Testing of Program Analyzers for Soundness and Precision Issues

David Kaindlstorfer (TU Wien, AT), Maria Christakis (TU Wien, AT), Anastasia Isychev (TU Wien, AT), and Valentin Wüstholz (Consensys – Wien, AT)

Main reference David Kaindlstorfer, Anastasia Isychev, Valentin Wüstholz, Maria Christakis: "Interrogation Testing of Program Analyzers for Soundness and Precision Issues", in Proc. of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 - November 1, 2024, pp. 319–330, ACM, 2024.
URL https://doi.org//10.1145/3691620.3695034

Program analyzers are critical in safeguarding software reliability. However, due to their inherent complexity, they are likely to contain bugs themselves, and the question of how to detect them arises. Existing approaches, primarily based on specification-based, differential, or metamorphic testing, have been successful in finding analyzer bugs, but also come with certain limitations. In this paper, we present interrogation testing, a novel testing methodology for program analyzers, to address limitations in existing metamorphic-testing techniques. Specifically, interrogation testing introduces two key innovations by (1) incorporating more information from analyzer queries to construct more powerful oracles, and (2) introducing a knowledge base that maintains a history of diverse queries. We implemented interrogation testing in Sher- lock and tested 8 mature analyzers—including model checkers, abstract interpreters, and symbolic-execution engines—that can prove the safety of assertions in C/C++ programs. We found 24 unique issues in these analyzers, 16 of which are soundness related, i.e., an analyzer does not report an assertion that can be violated. Our experimental evaluation demonstrates Sherlock's effectiveness by finding issues between 7x and 906x faster than our baseline, which is inspired by the state of the art.

3.7 UBGen: Generating UB Programs for testing Sanitizers

Shaohua Li (The Chinese University of Hong Kong, HK)

Joint work of Shaohua Li, Zhendong Su
Main reference Shaohua Li, Zhendong Su: "UBFuzz: Finding Bugs in Sanitizer Implementations", in Proc. of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024, pp. 435–449, ACM, 2024.
URL https://doi.org//10.1145/3617232.3624874

In this talk, I will introduce our new program generator UBGen, which can generate C programs with various undefined behaviors, such as buffer overflows and null pointer dereferences. We have used UBGen to fuzz sanitizers (ASan, UBSan, and MSan) and have successfully detected more than 10 false negative bugs, where sanitizers failed to report the UB in programs.

3.8 The Mopsa static analysis platform, and our quest to ease implementation & maintenance

Raphaël Monat (INRIA Lille, FR)

Joint work of Raphaël Monat, Abdelraouf Ouadjaout, Antoine Miné
Main reference Raphaël Monat, Abdelraouf Ouadjaout, Antoine Miné: "Easing Maintenance of Academic Static Analyzers", International Journal on Software Tools for Technology Transfer, Vol. CSV 2024 Special Issue, Springer Verlag, 2025.
URL https://doi.org//10.1007/s10009-024-00770-1

Mopsa is a Modular Open Platform for Static Analysis, whose goal is to encourage research and education in static analysis, by providing a fully-featured and extensible open-source platform and usable analyses built with it. In particular, analyses in Mopsa can reach a high expressivity, thanks to a framework allowing an extensive use of relational domains, which are able to infer linear constraints between variables. In this talk, we will see a brief overview of Mopsa, our main design decisions and the results we have able to obtain so far.

Implementations of static analyzers are time-consuming to develop and to maintain, but necessary to enable building further research upon the implementation. This talk will present the tools and techniques we have come up with to simplify the maintenance of Mopsa. First, we describe an automated way to measure precision that does not require any manual inspection of the results, improves transparency of the analysis, and helps discovering regressions during continuous integration. Second, we have taken inspiration from standard tools observing the concrete execution of a program to design custom tools observing the abstract execution of the analyzed program itself, such as abstract debuggers and profilers. Finally, we report on some cases of automated testcase reduction.

3.9 Semantic Metamorphic Testing for Finding Bugs in SMT Solvers.

Hakjoo Oh (Korea University – Seoul, KR)

Ensuring the correctness of SMT solvers is thus critical, as they serve as the cornerstone of a wide range of software engineering applications, from symbolic execution and program verification to program synthesis and repair. However, existing testing techniques, such as differential and metamorphic testing, have limitations: the former requires multiple solvers and is confined to shared functionality, while the latter is often restricted to simple, syntactic-preserving transformations. In this talk, I present DIVER, a technique that overcomes these limitations through semantic metamorphic testing. Unlike prior approaches, DIVER performs oracle-guided, unrestricted random mutations based on the semantic model of a formula, enabling it to uncover deep, solver-specific soundness and model-generation bugs that are out of reach for existing tools. Using DIVER, we have discovered 25 new bugs in Z3, CVC5, and dReal, including subtle logical errors that had persisted in production releases for years.

3.10 Testing Quantum Computing Platforms

Michael Pradel (Universität Stuttgart, DE)

Joint work of Michael Pradel, Matteo Paltenghi
Main reference Matteo Paltenghi, Michael Pradel: "MorphQ: Metamorphic Testing of the Qiskit Quantum Computing Platform", in Proc. of the 45th IEEE/ACM International Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023, pp. 2413–2424, IEEE, 2023.
URL https://doi.org//10.1109/ICSE48619.2023.00202

Quantum computing is a rapidly evolving field with the potential to revolutionize a wide range of industries. At the core of this revolution are quantum computing platforms, which – similar to traditional analyzers and compilers – analyze, optimize, and translate quantum programs. Unfortunately, like any complex software, these platforms are susceptible to bugs that can undermine the correctness and reliability of quantum applications. This talk presents two automated testing techniques designed to address this challenge. The first, MorphQ, is a metamorphic testing approach tailored to quantum computing platforms. It combines a generator of quantum input programs with a suite of program transformations that exploit quantum-specific metamorphic relationships to uncover inconsistencies. The second technique, QITE, introduces a cross-platform testing framework for quantum computing. It is based on a novel ITE process that generates equivalent quantum programs by iteratively (I) Importing assembly code into platform-specific representations, (T) Transforming the programs via platform-specific optimizations and gate conversions, and (E) Exporting them back to assembly. Both approaches have successfully identified a range of previously unknown bugs in widely used quantum computing platforms, including Qiskit, Pennylane, and Pytket, thereby contributing to the robustness and trustworthiness of this emerging field.

3.11 hevm, a flexible symbolic execution framework to verify EVM bytecode

Mate Soos (Ethereum – Berlin, DE)

Joint work of Mate Soos, dxo, Zoe Paraskevopoulou
Main reference Dxo, Mate Soos, Zoe Paraskevopoulou, Martin Lundfall, Mikael Brockman: "Hevm, a Fast Symbolic Execution Framework for EVM Bytecode", in Proc. of the Computer Aided Verification - 36th International Conference, CAV 2024, Montreal, QC, Canada, July 24-27, 2024, Proceedings, Part I, Lecture Notes in Computer Science, Vol. 14681, pp. 453–465, Springer, 2024.
URL https://doi.org//10.1007/978-3-031-65627-9_22

We present hevm, a symbolic execution engine for the EVM. hevm can prove safety properties for EVM bytecode or verify semantic equivalence between two bytecode objects. It exposes a user-friendly API in Solidity that allows end-users to define symbolic tests using almost the same syntax as they would for their usual unit tests. We evaluate our framework against state-of-the-art tools, using a comprehensive set of benchmarks. Our empirical findings demonstrate that hevm outperforms its counterparts, effectively solving a greater number of problems within competitive time frames.

3.12 Synthesizing Test Cases for Testing Type Checkers

Thodoris Sotiropoulos (ETH Zürich, CH)

Main reference Thodoris Sotiropoulos, Stefanos Chaliasos, Zhendong Su: "API-Driven Program Synthesis for Testing Static Typing Implementations", Proc. ACM Program. Lang., Vol. 8(POPL), pp. 1850–1881, 2024.
URL https://doi.org//10.1145/3632904

Type checkers are the most widely used form of static analysis, helping us identify bugs in programs during development. However, bugs in type checkers themselves can harm the programmer experience and, more critically, pose security risks, especially when they compromise the soundness of the checker.

In this talk, we address one of the central challenges in program generation for testing type checkers: saturation. We introduce THALIA, a framework that leverages APIs from real-world software libraries to synthesize small client programs designed to stress-test type checker implementations. The strength of THALIA comes from the inherent complexity of modern APIs, which often depend on advanced typing features such as parametric polymorphism and overloading. By exploiting these APIs, THALIA produces programs that exercise sophisticated typing behaviors without the need to explicitly generate those features from scratch.

When applied to popular type checkers, THALIA uncovered dozens of previously unknown bugs, many of which had been missed by existing program generation techniques.

References

1 Thodoris Sotiropoulos, Stefanos Chaliasos, Zhendong Su: API-Driven Program Synthesis for Testing Static Typing Implementations. Proc. ACM Program. Lang. 8(POPL): 1850-1881 (2024)
2 Stefanos Chaliasos, Thodoris Sotiropoulos, Diomidis Spinellis, Arthur Gervais, Benjamin Livshits, Dimitris Mitropoulos: Finding typing compiler bugs. PLDI 2022: 183-198
3 Stefanos Chaliasos, Thodoris Sotiropoulos, Georgios-Petros Drosos, Charalambos Mitropoulos, Dimitris Mitropoulos, Diomidis Spinellis: Well-typed programs can go wrong: a study of typing-related bugs in JVM compilers. Proc. ACM Program. Lang. 5(OOPSLA): 1-30 (2021)

3.13 Termination (Resilience) Analysis, and Bugs in Its Implementation

Caterina Urban (INRIA & ENS Paris, FR)

Joint work of Caterina Urban, Naïm Moussaoui Remil

We present a novel abstract interpretation-based static analysis for proving Termination Resilience, the absence of Robust Non-Termination vulnerabilities in software programs. Robust Non-Termination characterizes programs where an externally-controlled input can force infinite execution, independently of other uncontrolled variables. The approach is implemented in the open-source tool FuncTion. We conclude with an overview of the bugs that we accidentally found during its development, longing for a more principled way to uncover such issues.

3.14 Fuzzing Zero-Knowledge Infrastructure

Valentin Wüstholz (Consensys – Wien, AT), Maria Christakis (TU Wien, AT), Anastasia Isychev (TU Wien, AT)

Joint work of Christoph Hochrainer, Anastasia Isychev, Valentin Wüstholz, Maria Christakis
Main reference Christoph Hochrainer, Anastasia Isychev, Valentin Wüstholz, Maria Christakis: "Fuzzing Processing Pipelines for Zero-Knowledge Circuits", CoRR, Vol. abs/2411.02077, 2024.
URL https://doi.org//10.48550/ARXIV.2411.02077

Zero-knowledge (ZK) infrastructure is highly complex and highly critical for the correct operation of several privacy-focused applications, such as online voting and blockchains; that is, a single bug can result in massive financial and reputational damage. To find such potential million-dollar bugs before they are exploited, we have developed a novel fuzzing technique that can find logic flaws that impact soundness or completeness of ZK infrastructure. Our fuzzer has already found 20 such issues in four ZK systems, namely Circom, Corset, Gnark, and Noir.

3.15 On Test Oracles for Program Analyzers and Verifiers

Chengyu Zhang (Loughborough University, GB)

Joint work of Chengyu Zhang, Zhendong Su, Dominik Winterer, Ting Su, Geguang Pu, Fuyuan Zhang, Yichen Yan, Weigang He, Peng Di, Mengli Ming, Shijie Li, Yulei Sui
Main reference Chengyu Zhang, Ting Su, Yichen Yan, Fuyuan Zhang, Geguang Pu, Zhendong Su: "Finding and understanding bugs in software model checkers", in Proc. of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019, pp. 763–773, ACM, 2019.
URL https://doi.org//10.1145/3338906.3338932

Program analyzers and verifiers are fundamental to building reliable software systems. Consequently, developing effective methodologies and practical tools to solidify these foundational components is critical. However, constructing test oracles for program analyzers and verifiers poses significant challenges due to the inherent complexity of their tasks.

In this talk, I will summarize three effective methodologies for building test oracles for program analyzers, verifiers, and their underlying tools. These methodologies have proven successful in uncovering thousands of bugs across various software, including SMT solvers, software and hardware model checkers, program verifiers, and static analyzers.

The talk will focus on both the theoretical and practical challenges of constructing test oracles for program analyzers and verifiers, and introduce the latest advances.

References

1 Chengyu Zhang, Ting Su, Yichen Yan, Fuyuan Zhang, Geguang Pu, Zhendong Su: Finding and understanding bugs in software model checkers. ESEC/SIGSOFT FSE 2019: 763-773
2 Dominik Winterer, Chengyu Zhang, Zhendong Su: Validating SMT solvers via semantic fusion. PLDI 2020: 718-730
3 Dominik Winterer, Chengyu Zhang, Zhendong Su: On the unusual effectiveness of type-aware operator mutations for testing SMT solvers. Proc. ACM Program. Lang. 4(OOPSLA): 193:1-193:25 (2020)

4 Weigang He, Peng Di, Mengli Ming, Chengyu Zhang, Ting Su, Shijie Li, Yulei Sui: Finding and Understanding Defects in Static Analyzers by Constructing Automated Oracles. Proc. ACM Softw. Eng. 1(FSE): 1656-1678 (2024)
5 Chengyu Zhang, Zhendong Su: SMT2Test: From SMT Formulas to Effective Test Cases. Proc. ACM Program. Lang. 8(OOPSLA2): 222-245 (2024)

4 Panel discussions

4.1 Proactively Testing Program Analyzers and Verifiers: Is It Worth It?

Thodoris Sotiropoulos (ETH Zürich, CH)

The first discussion panel of the seminar explored whether it is worthwhile to proactively test program analyzers and verifiers, particularly through techniques such as fuzzing. Arguments were presented on both sides. We clarify that these arguments do not suggest that program analyzers and verifiers do not matter. Instead, we question whether it is important to *proactively* test them using fuzzing to find edge case bugs.

Arguments against proactive testing.

- Software systems rely on many components including user code (the code that should be verified or analyzed), specifications (describing what should be checked), libraries, compilers, operating systems, testing infrastructure and test suites, and program analyzers and verifiers. Why should our attention specifically go to program analyzers and verifiers? Isn't it more likely there will be serious programs with specifications, or issues in library code, or compiler bugs?
- For a verifier bug to cause a severe failure, several conditions must align:
  - A bug exists in the analyzer.
  - A bug exists in the user code.
  - The analyzer bug masks the user bug.
  - The masked bug is not caught by other means (testing, reviews, etc.).
  - The masked bug leads to a serious real-world failure.

  The probability of all these factors aligning was considered small. As a concrete example, a bug in Dafny was mentioned, which was deemed too much of an edge case to realistically occur in practice.
- Proactive fuzzing requires ongoing investment in maintaining target systems across versions, which may not always be feasible or cost-effective.
- Even when bugs are found, developers may not respond effectively, whether because they are overwhelmed or because they prioritize other issues. This raises concerns about the return on investment in proactive testing program analyzers and verifiers.

Arguments in favor of proactive testing.

- A single fault in a verifier can affect many users and programs at the same time, amplifying its potential consequences compared to faults in user-specific code.
- If analyzers and verifiers are perceived as untrustworthy, the entire software ecosystem suffers. For example, large-scale systems like Amazon's infrastructure, which processes billions of queries daily, depend critically on trustworthy analyzers.
- As software development increasingly moves toward specification-driven approaches (programs verified rather than tested), the correctness of analyzers and verifiers becomes central to reliability.

- Bugs that affect users cannot always be predicted in advance. Proactive testing provides a way to uncover such issues before they manifest in practice.
- Past experience with unreliable compilers demonstrates the importance of testing infrastructure tools early, rather than assuming they are inherently reliable.
- Proactive testing adds an important safeguard. Even if most testing is carried out by users, regression testing and fuzzing provide an early line of defense.

4.2 Fault De-duplication, Prioritization, And Other Considerations

Thodoris Sotiropoulos (ETH Zürich, CH)

This session discussed several topics proposed by the attendees during the seminar. The discussions were around the importance of faults in program analyzers/verifiers, bug de-duplication and prioritization, saturation of program generation, AI for testing, and performance testing of program analyzers/verifiers.

Fault importance and prioritization. A key theme was the prioritization of bugs. Participants distinguished between bug de-duplication (eliminating duplicate reports) and bug prioritization (deciding which unique bugs are most important to address). Many attendees emphasized that de-duplication based on crash signatures can be misleading, as many similar crashes may stem from distinct underlying issues. Ultimately, only the tool developers are in a strong position to correctly identify duplicates.

The type of discovered issues may also influence prioritization. For example, some attendees shared their experience with their interaction with the Kotlin development team. In particular, Kotlin developers recently focused on regressions between Kotlin v1.0 and v2.0, rather than on bugs present in both versions, even when those bugs affected soundness of type checkers.

AI for testing. The panel discussed AI-assisted testing, including the Fuz4all project [1]. While details were not extensively covered, the discussion reflected growing interest in leveraging machine learning and AI to improve bug-finding effectiveness.

Saturation of program generation. The notion of saturation, that is, the point at which a program generator appears to stop finding new bugs, was revisited. Attendees argued that saturation is not necessarily negative: once saturation is reached, any additional bugs discovered are highly likely to be novel and meaningful. However, other attendees advocated that "there is no end": even slight modifications to a generator or support for new features often lead to the discovery of new classes of bugs.

Performance testing. Performance was briefly mentioned, with particular focus on performance-critical software such as SMT solvers. Since analyzers often depend on solvers, small regressions in solver speed or behavior can have side-effects in program analysis tools. Sometimes, this can cause analyses to fail unexpectedly. While this was only lightly discussed, the brittleness of solver performance was recognized as a practical concern.

References

1 Chunqiu Steven Xia, Matteo Paltenghi, Jia Le Tian, Michael Pradel, Lingming Zhang: Fuzz4All: Universal Fuzzing with Large Language Models. ICSE 2024: 126:1-126:13

Participants

- Cristian Cadar
Imperial College London, GB
- Maria Christakis
TU Wien, AT
- Pascal Cuoq
TurstInSoft – Paris, FR
- Alastair F. Donaldson
Imperial College London, GB
- Karine Even-Mendoza
King's College London, GB
- Markus Fleischmann
TU Wien, AT
- Amber Gorzynski
Imperial College London, GB
- Manuel Hermenegildo
IMDEA Software Institute – Pozuelo de Alarcón, ES & UPM – Madrid, ES
- Anastasia Isychev
TU Wien, AT
- David Kaindlstorfer
TU Wien, AT
- Shaohua Li
The Chinese University of Hong Kong, HK
- Muhammad Numair Mansur
Amazon Web Services – Berlin, DE
- Raphaël Monat
INRIA Lille, FR
- Hakjoo Oh
Korea University – Seoul, KR
- Michael Pradel
Universität Stuttgart, DE
- John Regehr
University of Utah – Salt Lake City, US
- Mate Soos
Ethereum – Berlin, DE
- Thodoris Sotiropoulos
ETH Zürich, CH
- Hao Sun
ETH Zürich, CH
- Caterina Urban
INRIA & ENS Paris, FR
- Valentin Wüstholz
Consensys – Wien, AT
- Chengyu Zhang
Loughborough University, GB

Report from Dagstuhl Seminar 25261

Future of Human-Centered Privacy

Zinaida Benenson[*1], Simone Fischer-Hübner[*2], Heather Richter Lipford[*3], and William Seymour[*4]

1 **Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), DE.** zinaida.benenson@fau.de
2 **Karlstad University, Chalmers University of Technology & University of Gothenburg, SE.** simone.fischer-huebner@kau.se
3 **University of North Carolina at Charlotte, US.** heather.lipford@uncc.edu
4 **King's College London, UK.** william.seymour@kcl.ac.uk

Abstract

The Dagstuhl Seminar on The Future of Human-Centered Privacy (25261), held from June 22–27, 2025, brought together researchers from academia and industry to discuss key issues at the intersection of privacy and human-computer interaction (HCI) research. This article summarizes the main discussion topics, and presents the summary of the outputs of five working groups that discussed: i) Measurement, Methods, and Ethics; ii) Supporting Developers; iii) AI for Privacy/Privacy for AI; iv) Consent, Control, and Communication; and v) Collective Privacy. This seminar was a continuation of a previous seminar held at King's College London on June 5–7, 2023 which laid the groundwork for the present seminar through it's discussion on the topics of inclusive privacy, multiuser privacy, privacy and AI, and privacy communication.

Seminar June 22–27, 2025 – https://www.dagstuhl.de/25261
2012 ACM Subject Classification Security and privacy → Human and societal aspects of security and privacy
Keywords and phrases Privacy, Human-computer Interaction, AI
Digital Object Identifier 10.4230/DagRep.15.6.84

1 Executive Summary

Zinaida Benenson (Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), DE)
Simone Fischer-Hübner (Karlstad University, Chalmers University of Technology & University of Gothenburg, SE)
Heather Richter Lipford (University of North Carolina at Charlotte, US)
William Seymour (King's College London, UK)

Human-centered privacy resides in the intersection of privacy and human-computer interaction (HCI) research. It investigates users' privacy perceptions, concerns, and awareness in various settings, and also the understanding, usefulness, and usage of various privacy-enhancing technologies. On the one hand, the advance of Internet of Things, smart spaces, and AI have raised new questions that need to be investigated, e.g., how to negotiate privacy settings in the presence of different users of the same system, or how to improve the transparency of AI systems. On the other hand, there are many questions that have been explored for decades, but need to be adapted to these new areas and domains, such as "What is a privacy decision?" and "What information do users need to make a privacy decision?". Moreover, the multitude

[*] Editor / Organizer

of users also includes at-risk and vulnerable populations that interact (sometimes unwillingly or unknowingly) with digital systems, and require additional research to understand their needs.

This seminar brought together academic and industry experts from a range of disciplines to discuss these issues, which had been seeded at a preceding workshop hosted at King's College London in 2023.

Seminar participants initially convened for an opening session with the project organisers that included a short round of introductions with sharing of current research interests by all participants. Following this, there was a talk from Sören Preibusch with a practical perspective on human-centred privacy and a panel debate on Privacy in the age of AI. During the short introductory presentation sessions, invited talk, and the panel sessions, seminar participants noted key human-centred privacy aspects of interest and in need of future research. These notes were then clustered by a small group of participants into ten different themes. After voting and discussion, working groups were formed to elaborate on the following five:

- Measurement, Methods, and Ethics
- Supporting Developers
- AI for Privacy, Privacy for AI
- Consent, Control, and Communication
- Collective Privacy

These working groups met throughout the remaining duration of the seminar, periodically reporting back to the main group. The outputs of the five groups form the main body of this seminar report, and include an overview of the main open problems in the area, an overview of current approaches, and promising directions for future work. These take the form of key research questions, potential solutions in need of study, and roadmaps for the development of research capacity.

To round up the seminar, all working groups presented their future work directions on the last day. Afterwards, we discussed future intergroup activities: closed workshops, open workshops (e.g., at SOUPS or CHI conferences), joint projects and publications, possible funding (e.g., EU Cost Actions, Erasmus+) and research exchange visits at various universities.

2 Table of Contents

3 Overview of Talks and Panels

3.1 Human-centered Privacy in Practice

Sören Preibusch (Bundesinstitut für Risikobewertung – Berlin, DE, soeren.preibusch@bfr.bund.de)

In my talk, I deliver a practical perspective on human-centred privacy, drawing on my experience in a large tech company, a data protection supervisory authority, and a federal agency where I'm currently the head of IT. Whereas the goals in each of the roles vary, there is often both an operational and a strategic aspect to the work and we must consider three questions when building for privacy: What are we optimising for (e.g., compliance, improved experiences)? Whom are we designing for (e.g., employees, citizens, users)? Who is the adversary (e.g., external attackers, insider threats)?

I conclude that the practical effort in achieving good privacy arises both at the product-level (to be solved through managerial decisions) and at the programme-level (to be solved through capacity building).

I encourage our seminar to embark on research that tackles two challenges: First, how might we focus human labour on the most value-added activities – and automate the rest? Second, how might we do privacy right when it's not the primary task?

3.2 Panel Discussion on "Privacy in the age of AI"

Bettina Berendt (Weizenbaum Institute, TU Berlin, DE & KU Leuven, BE, berendt@tu-berlin.de)
Marc Langheinrich (Università della Svizzera italiana – Lugano, CH, marc.langheinrich@usi.ch)
Cristiana Santos (School of Law, Utrecht University, NL, c.teixeirasantos@uu.nl)

3.2.1 Statements

Marc Langheinrich

There will be more AI-generated attacks and scams which we will need to address. There is also the potential for AI to be used as personalized privacy/security advisors to keep users informed of these risks. For example, AI could be used to give people a better idea of what the consequences and inferences of their data disclosures.

Cristiana Santos

Developers need to make privacy-related decisions without sufficient legal guidance. They may now use LLMs, whose answers can be generic, lack legal relevance, and be unreliable. Thus, the community needs to help developers to not overly rely on AI.

Bettina Berendt

Thesis 1: We need to consider both "users" and "privacy" in a larger sense if we want to explore "the future": (a) from users to stakeholders or "affected persons" (as in the AI Act); (b) from privacy as confidentiality or control to "the freedom from unreasonable constraints on the construction of one's identity" – and maybe move beyond this focus on identity.

Thesis 2: Right now, we (whether as communities, countries, or humankind) arguably have bigger fish to fry than – to be polemic – avoiding unwanted ads: 1. at least putting restrictions on the ever-increasing concentration of economic (and with it political) power, a concentration much furthered by AI; 2. digital sovereignty, including cybersecurity for critical infrastructures; 3. the survival of the planet; 4. (re)instating the rule of law, internally, in international relations, and with a view to human rights. The bad news is that threats to all four appear to be spurred on by AI.

Thesis 3: The good news is that the protection of personal data, in particular via the core principle of data minimisation, can serve as a starting point in the fight for 1., 2. and 3. As regards 4., this probably exceeds the scope of our discussion here ...

3.2.2 Discussion

The seminar participants discussed the challenge of defining privacy when an AI model is the potential adversary. This may lead to different or new interpretations of privacy. Who and how we address these challenges will depend on how stakeholders define privacy.

4 Working Group Reports

4.1 Measurement, Methods, and Ethics

Cori Faklaris (University of North Carolina – Charlotte, US, cfaklari@charlotte.edu)
Yixin Zou (Max Planck Inst. for Security and Privacy – Bochum, DE, yixin.zou@mpi-sp.org)
Adam Jenkins (King's College – London, UK, adam.jenkins@kcl.ac.uk)
Anastasia Sergeeva (University of Luxembourg, LU, anastasia.sergeeva@uni.lu)
Apu Kapadia (Indiana University Bloomington, US, kapadia@iu.edu)
Daricia Wilkinson (Arizona State University – Tempe, US, daricia.wilkinson@asu.edu)
Sameer Patil (University of Utah – Salt Lake City, US, sameer.patil@utah.edu)
Simran Munot (Max Planck Institute for Informatics, DE, smunot@mpi-inf.mpg.de)

4.1.1 Introduction

We explored the challenges related to methodological rigor, participant consent, the validity of new data sources, and the ethical boundaries of AI-generated insights. Our discussions also covered foundational issues in research, including the reporting of demographics, the evolution of privacy education, and the persistent gaps in how we measure and evaluate privacy itself. This summary outlines the key problems, the current state of the art, open research questions, and a potential roadmap for the HCP research community.

4.1.2 Problems and Challenges

AI-Mediated Research. The use of AI tools, such as large language models (LLMs) for thematic analysis or AI companions for transcription, introduces questions of rigor, bias, and consent. There is significant scepticism about whether participants can genuinely consent to AI-mediated processes they may not fully understand and whether these methods can preserve the empathetic, nuanced core of qualitative inquiry. Furthermore, publishing findings based on AI-generated or synthetic data carries risks to authenticity and may undermine trust in research outcomes.

Synthetic Data. While industry interest in using synthetic data for UX research is growing, this approach risks undermining the authenticity and empathy that are central to qualitative research. Synthetic data may flatten or misrepresent the nuanced experiences of real users, and its use as a privacy-enhancing technology is being questioned, as it may not be immune to linkage attacks or attribute inference.

Demographics Reporting. The HCP field lacks common guidelines for collecting and reporting demographic data. This is complicated by GDPR constraints that mandate data minimization, which can conflict with the goal of comprehensive demographic reporting for transparency and bias detection. Moreover, first-order attributes such as age and gender are often coarse proxies for the lived experiences (e.g., discrimination) that are truly relevant.

Measurement and Evaluation Gaps. Many established scales and measurements for privacy concepts (e.g., privacy concern) are being challenged. Recent work shows that participants often interpret survey items differently than researchers intend, questioning the validity of these instruments. The rapid evolution of AI is likely further changing privacy expectations and behaviors, making our current measurement tools potentially obsolete.

Education and Curriculum. The integration of AI into education fosters an over-reliance on tools that can diminish critical thinking and problem-solving skills. It also introduces new avenues for academic dishonesty and erodes the essential human interaction in learning. For privacy education specifically, there is an urgent need to update curricula to address the complexities and interplay of AI, data transparency, and the often-flawed user mental models of how these systems operate.

4.1.3 State of the Art

The current state of HCP is similar to that of computing as a whole. It is characterized by rapid technological advancement running ahead of methodological and ethical consensus. Notably, the **use of AI in research** is seen as inevitable and is already widespread, with researchers using it for faster data processing and code development. However, studies comparing human and AI coders on qualitative data to date show significant discrepancies and only moderate overlap, highlighting reliability issues. Similarly, the **use of synthetic data** has been proposed for decades and applied in fields like medicine and political science. However, recent studies in HCI show that LLM-generated data, while plausible, is less diverse, prone to factual errors, and can contain biases (e.g., related to age) compared with human-generated data.

While the HCP field lacks its own standards for **demographics reporting**, other fields such as computing education and clinical research have established guidelines that standardize such publications and make them easily comparable. Specific guidelines are also available for asking about sensitive dimensions such as gender and ethnicity. However, broader societal shifts in understanding identity may require these to be updated.

A significant body of work exists on **measuring privacy concern and related constructs**, originating from law, public policy, and information systems. However, recent critiques using techniques like corpus linguistics have challenged the construct validity of many widely used scales, revealing a disconnect between researcher intent and participant interpretation.

AI in Education (AIED) is increasingly prevalent for its potential to enhance teaching and learning. However, the focus has often been on the technological implementation, with a growing body of work now highlighting the negative impacts on human cognition, critical thinking, and academic integrity.

4.1.4 Key Research Questions

To move forward, our community must address several key research questions:

1. **AI-Mediated Research & Synthetic Data:**
   - Under what circumstances is it feasible and ethical to use synthetic data or AI-assisted analysis in HCP research?
   - How can we develop benchmarks to validate the quality, representativeness, and authenticity of synthetic data?
   - How can we design and implement practices that embed meaningful transparency and trust-building into AI-mediated research, ensuring genuine participant consent?
   - How can we best collect, de-identify, and disseminate inclusive datasets for training AI models to offset dominant Western-focused cultural biases and amplify marginalized voices?
2. **Demographics and Measurement:**
   - How can we, as a community, develop guidelines to balance the goals of scientific transparency and rigor with the ethical needs for participant privacy and data minimization, especially under constraints like GDPR?
   - Moving beyond coarse proxies such as age and race, which are second- and third-order attributes (e.g., experienced discrimination, technical literacy) should we focus on to better understand our participants and their contexts?
   - How can we develop and validate new measurement instruments for privacy constructs that are robust to misinterpretation and reflect contemporary understandings of privacy in an AI-driven world?
3. **Education:**
   - How should we redesign HCP curricula to teach students not just how to *use* AI tools, but how to *critically reflect* on their societal, ethical, and personal impact?
   - What competency frameworks are needed to define what students should know (awareness) and be able to do (proficiency) regarding AI and privacy?

4.1.5 Solution Ideas and Directions

Addressing these questions requires a multi-faceted approach combining technical, methodological, and community-driven efforts.

- *Develop Community Guidelines and Standards:* The HCP community should work towards shared guidelines for the ethical use of AI in research, the reporting of demographics, and the validation of synthetic data. This could involve requiring explicit discussions of these considerations in paper submissions and review forms.

- *Adopt a "Human-in-the-Loop" Ethos:* For both research and education, we must emphasize critical human oversight. In research, this means not blindly accepting AI-generated analysis but using it as a tool to augment human intellect. In education, this means designing assignments where students must reflect on and critique the AI's output, with the student's critical judgment being the final word.
- *Create and Validate New Instruments and Taxonomies:* We need to develop and empirically validate new survey instruments for privacy. Additionally, creating a taxonomy of first-, second-, and third-order demographic attributes could help researchers make more intentional choices about data collection.
- *Establish Educational Resources:* There is an urgent need for repositories of syllabi, course resources, and hands-on learning modules for HCP education. These resources should be modelled on successful initiatives in related fields (e.g., usable security) and should explicitly address AI ethics and critical reflection.
- *Promote Transparency:* Researchers should be transparent about their methods, including documenting when and how AI was used. [Ex: The first author used Gemini Pro and Copilot for Education for helping to condense and aggregate the working group's documents.] For demographics, this includes reflecting on why certain attributes were collected and others were not. For education, this may mean making it a rule that students must disclose their use of AI.

To carry out this work will require resources for supporting the actual research and for developing and sustaining the types of people required. It will involve interdisciplinary collaboration between HCI researchers, ethicists, legal scholars (especially those specializing in data protection), data scientists, and educators. Skills in psychometrics, corpus linguistics, and qualitative methods are essential for developing and validating new measurements and analysis techniques. Furthermore, the HCP community needs funding to support the creation of inclusive datasets, the development of new measurement tools, and the organization of community-building events. We also need robust, shared infrastructure, such as secure repositories for data and educational materials.

Finally, we encourage recognizing the development of datasets, tools, and curricula as valuable academic contributions (on par with traditional publications). This will align incentives with the outlined agenda and with their importance for actually being able to create a more secure, private, and trustworthy future of computing.

4.1.6 Roadmap Development

Towards addressing the above concerns, we envision a progression of research and community action over the next decade.

Immediate (Next 1-2 Years).

- **Initiate Dialogue:** Form working groups and organize workshops at major HCI conferences (e.g., CHI, SOUPS, CSCW) dedicated to creating draft guidelines for AI in research and demographic reporting.
- **Curriculum Redesign:** Begin redesigning individual courses to incorporate critical AI reflection and hands-on learning modules. Start building an ad-hoc, shared repository for syllabi and course materials.
- **Incorporate into CFPs:** Program committees for major venues can begin to encourage or require reflection statements on the use of AI and demographic reporting choices in submissions.

Next Three Years.

- **Publish Community Guidelines:** Formalize and publish initial versions of community guidelines for AI use and demographic reporting.
- **Develop Competency Frameworks:** Establish a clear competency framework for AI and privacy education, defining core knowledge and skills.
- **Validate New Measures:** Conduct large-scale validation studies of new and existing privacy measurement instruments to establish a new set of reliable tools for the community.

Next Decade.

- **Establish Accreditation and Standards:** Work towards formal accreditation for educational programs that teach AI and privacy, similar to standards in cybersecurity.
- **Build Robust Infrastructure:** Develop and maintain community-wide infrastructure, including repositories for inclusive, de-identified datasets for training AI models, and platforms for sharing validated educational resources.
- **Longitudinal Studies:** Conduct long-term research tracking how technological affordances (like human-like chatbots) impact privacy behaviors and how societal understandings of privacy evolve over time.

References

1. Abdulrahman M. Al-Zahrani and Talal M. Alasmari. 2024. Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications. Humanities & social sciences communications 11, 1: 1–12. https://doi.org/10.1057/s41599-024-03432-4
2. Michael Benisch, Patrick Gage Kelley, Norman Sadeh, and Lorrie Faith Cranor. 2011. Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs. Personal and Ubiquitous Computing 15, 7: 679–694. https://doi.org/10.1007/s00779-010-0346-0
3. Jessica Colnago, Lorrie Faith Cranor, Alessandro Acquisti, and Kate Hazel Stanton. 2022. Is it a concern or a preference? An investigation into the ability of privacy scales to capture and distinguish granular privacy constructs. and Security (SOUPS 2022). Retrieved from https://www.usenix.org/conference/soups2022/presentation/colnago
4. Serge Egelman and Eyal Peer. 2015. Predicting Privacy and Security Attitudes. SIGCAS Comput. Soc. 45, 1: 22–28. https://doi.org/10.1145/2738210.2738215
5. Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 433, 1–19. https://doi.org/10.1145/3544548.3580688
6. Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. LLMs as research tools: A large scale survey of researchers' usage and perceptions. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2411.05025
7. Abdul Majeed. 2023. Attribute-centric and synthetic data based privacy preserving methods: A systematic review. Journal of Cybersecurity and Privacy 3, 3: 638–661. https://doi.org/10.3390/jcp3030030
8. Naresh K. Malhotra, Sung S. Kim, and James Agarwal. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. Information Systems Research 15, 4: 336–355. https://doi.org/10.1287/isre.1040.0032

9 Sameer Patil, Greg Norcie, Apu Kapadia, and Adam J. Lee. 2012. Reasons, rewards, regrets: privacy considerations in location sharing as an interactive practice. In Proceedings of the Eighth Symposium on Usable Privacy and Security. https://doi.org/10.1145/2335356.2335363

10 Sören Preibusch. 2013. Guide to Measuring Privacy Concern: Review of Survey and Observational Instruments. Int. J. Hum. -Comput. Stud. 71, 12: 1133–1143. https://doi.org/10.1016/j.ijhcs.2013.09.002

11 Anastasia Sergeeva, Björn Rohles, Verena Distler, and Vincent Koenig. 2023. "We Need a Big Revolution in Email Advertising": Users' Perception of Persuasion in Permission-based Advertising Emails. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 652, 1–21. https://doi.org/10.1145/3544548.3581163

12 Sarah Spiekermann and Lorrie Faith Cranor. 2009. Engineering Privacy. IEEE transactions on software engineering 35, 1: 67–82. https://doi.org/10.1109/tse.2008.88

13 Sarah Tabassum, Nishka Mathew, and Cori Faklaris. 2025. Privacy on the Move: Understanding Educational Migrants' Social Media Practices through the Lens of Communication Privacy Management Theory. In Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS '25). Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3715335.3735453

14 Mohammad Tahaei, Adam Jenkins, Kami Vaniea, and Maria Wolters. 2021. "I don't know too much about it": On the security mindsets of Computer Science students. arXiv [cs.CR], 27–46. https://doi.org/10.1007/978-3-030-55958-8_2

15 Kurt Thomas, Patrick Gage Kelley, David Tao, Sarah Meiklejohn, Owen Vallis, Shunwen Tan, Blaž Bratanič, Felipe Tiengo Ferreira, Vijay Kumar Eranti, and Elie Bursztein. 2025. Supporting Human Raters with the Detection of Harmful Content using Large Language Models. 2772–2789. Retrieved from https://ieeexplore.ieee.org/abstract/document/11023319/

16 Richard Van Noorden and Jeffrey M. Perkel. 2023. AI and science: what 1,600 researchers think. Nature 621, 7980: 672–675. https://doi.org/10.1038/d41586-023-02980-0

17 Daricia Wilkinson, Moses Namara, Karla Badillo-Urquiola, Pamela J. Wisniewski, Bart P. Knijnenburg, Xinru Page, Eran Toch, and Jen Romano-Bergstrom. 2018. Moving Beyond a "one-size fits all": Exploring Individual Differences in Privacy. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, 1–8. Retrieved from https://dl.acm.org/doi/abs/10.1145/3170427.3170617

18 Pamela J. Wisniewski, Bart P. Knijnenburg, and Heather Richter Lipford. 2017. Making privacy personal: Profiling social network users to inform privacy education and nudging. International journal of human-computer studies 98: 95–108. https://doi.org/10.1016/j.ijhcs.2016.09.006

4.2 Supporting Developers

Sören Preibusch (BfR – Berlin, DE, soeren.preibusch@bfr.bund.de)
Nataliia Bielova (Inria – Rennes, FR, nataliia.bielova@inria.fr)
Dominik Herrmann (University of Bamberg, DE, dominik.herrmann@uni-bamberg.de)
Alena Naiakshina (University of Cologne, DE, alena.naiakshina@uni-koeln.de)
Cristiana Santos (School of Law, Utrecht University, NL, c.teixeirasantos@uu.nl)
Ha Dao (MPI-INF – Saarbrücken, DE, hadao@mpi-inf.mpg.de)
Bettina Berendt (Weizenbaum Institute, TU Berlin, DE & KU Leuven, BE, berendt@tu-berlin.de)

4.2.1 Introduction

Software products and services shape societies and peoples' lives. These software products don't magically appear – they are created by developers and then used by many more users, potentially by orders of magnitude more. Thus, the design and coding choices of developers determine the quality of software products – including the quality of privacy[1] implementations. This differentiates developers from consumers who usually only make privacy choices for themselves or a small number of others (see Chapter 4.5 on Collective Privacy). Unfortunately, many developers are ill-equipped to shoulder this broad impact: they lack both expertise and advice on privacy, and formal training in software engineering. In addition, being a developer can mean many different things, depending on the roles and the contexts: First, software development is never just about writing code. It encompasses the entire Software Development Lifecycle, from planning and design to deployment and maintenance, as well as all the internal processes, guidelines, and communication that enable and sustain that lifecycle. The resulting jobs to be done include writing code, testing, product and project management, requirements engineering and software architecting – each one associated with specific privacy questions and decisions [3]. In an ideal world, these jobs are distributed among experts who build specialized skills. In reality, governance structures differ and some jobs are combined, skipped, or executed by a "one-person band" [29].

Second, different contexts like the development of a new website or app, software maintenance, ad-hoc scripting to automate admin tasks, or database development all come with their own support needs. In each one of these contexts, developers routinely switch between producing code and consuming code (e.g., embedding third-party services or libraries, reusing code snippets).

4.2.2 Legal Background

This section defines the scope of relevant EU laws applicable to this report and offers conceptual legal clarifications, such as responsibility, accountability, data controller, data processor, liability.

General Data Protection Regulation (GDPR). The GDPR applies to the processing of personal data. Personal data refers to any information that renders a person identified or identifiable.

[1] We understand "privacy" as mediated via data protection.

Responsibility: Responsibility for compliance with GDPR obligations is determined based on the roles of the parties involved in the processing of personal data. The GDPR acknowledges the concept of "control" to assign responsibility for GDPR compliance (and administrative fines for non-compliance) to the data controller.

Controller: The controller is the entity that determines the "purposes" and "means" of processing personal data.

- Purposes: for example, developers determine the "purposes" of processing once they define the functionality needed for their own webpage/app service (either for compliance, site protection, user authentication, advertising, analytics, etc.).
- Means: concern more practical aspects of implementation, such as the choice of a particular type of hard- or software or the detailed security measures which may be left to the processor to decide on. Means are to be determined by the controller and processor. Developers determine the "means" when they embed that tool in its webpage service, thus triggering the start of the processing of personal data which would not be possible otherwise.
- Essential means: are traditionally and inherently reserved to the controller. "Essential means" are means that are closely linked to the purpose and the scope of the processing. Data controllers determine the "essential means" of processing by defining the type of data to be processed, the data recipients, data subjects and duration of processing. As such, a developer determines essential means (and it is thus a data controller) when they can terminate the processing, by simply removing the tool from its webpage.

The controller bears primary responsibility for compliance with GDPR principles. The controller is responsible for ensuring a legal basis for processing (e.g., consent), transparency and information provision, that data is processed for determined and specific purposes, data minimization, data protection by design and by default, responding to data subject rights, and adopting security measures. Case-law from the Court of Justice of the EU established that a given actor can be a controller if there is a purpose for data processing, even: (i) without having access to the data; (ii) without knowing that one processes personal data (not deliberately targeting personal data as such or wrongfully assessing that one does not process personal data); (iii) when choosing a third-party tool that allows the processing of personal data. These requirements for legal responsibility are important considering the challenges that several developers face, described in section 4.2.3, herein briefly referenced for clarity (in particular: privacy is not a priority for developers; developers do not feel responsible for privacy compliance; developers have difficulty understanding privacy requirements; developers depend on and trust third-party services; developers are manipulated by third-party tools).

Processor: The processor is the entity that processes data on behalf of the controller. The processor is responsible to follow only documented instructions from the controller, implement appropriate security, maintain records of processing, and notify data breaches to the controller.

Fines for GDPR infringements: When deciding on the amount of the administrative fine in each individual case, data protection authorities shall give due regard to the intentional or negligent character of the infringement; the nature, gravity and duration of the infringement taking into account the nature, scope or purpose of the processing concerned as well as the number of data subjects affected and the level of damage suffered by them. For especially severe violations, listed in Art. 83(5) GDPR, the fine can be up to 20 million euros, or in the case of an undertaking, up to 4% of their total global turnover of the preceding fiscal year, whichever is higher. But even the catalogue of less severe

violations in Art. 83(4) GDPR sets forth fines of up to 10 million euros, or, in the case of an undertaking, up to 2% of its entire global turnover of the preceding fiscal year, whichever is higher.

Liability for damages caused by a GDPR violation: A data subject has the right to compensation for material or non-material damage caused by a GDPR violation (Art. 82). A controller or processor is liable unless they can prove they were not responsible. Liability may be sole (if only one party was at fault), joint, or several (if multiple parties contributed).

Product Liability Directive (PLD). The PLD aims to protect consumers from defective products. It also covers digital products, such as software, digital manufacturing files, and AI systems placed on the market. Personal data is not a product under the PLD definition. The GDPR provides a dedicated liability framework for personal data.

Manufacturer: Manufacturers include both individuals and legal entities who develop, manufacture, or produce products, even if they create products solely for their own use. Additionally, anyone who designs a product or has it manufactured and then attaches their name to it ("quasi-manufacturer"), is also considered a manufacturer.

Liability for defective products: Manufacturers can be held liable if their products are defective. A product is deemed defective if it does not meet the expected or legally required level of safety. When assessing defectiveness, factors such as the product's presentation, characteristics, labelling, foreseeable use, and the impact of other products used in conjunction with it are taken into account (Art. 7). These requirements can stem from other EU laws, including the AI Act (AIA) or Cyber Resilience Act (CRA). The directive excludes pure violations of privacy.

4.2.3 Challenges Description

Basic setting: Developers' economic setting, attitudes, and knowledge. In this section, we examine how developers' economic context, organizational constraints, and knowledge limitations shape their attitudes and practices toward privacy.

- **Privacy is not a priority for developers.** Developers often do not prioritize privacy in their work. Prior studies have found that organizational dynamics can actively discourage developers from engaging with privacy. In particular, negative privacy climates – where privacy is viewed as a barrier or secondary concern – can shape developer behavior and expectations, disincentivizing attention to privacy concerns [10]. Beyond organizational culture, practical constraints such as limited budgets, tight deadlines, and competing feature demands further reduce the incentive to prioritize privacy during development [4]. Even among website owners, who may be considered a distinct subgroup of developers, privacy is not a consistent consideration. Their decisions are largely driven by business goals, personal motivations, and available resources, which often override concerns about data protection [29]. Notably, privacy rarely factors into their selection of third-party services, despite the privacy risks these integrations may pose [37].
- **Developers do not feel responsible for privacy compliance.** Developers often do not perceive themselves as responsible for ensuring the protection of privacy. Instead, privacy is frequently regarded as a legal matter, leading many developers to defer responsibility to legal or compliance teams. This perception creates a sense of limited control and can result in frustration when developers are required to engage with legal frameworks or implement regulatory requirements they do not feel empowered to influence [12]. Moreover, responsibility for privacy is often diffuse within organizations, so developers

may not know whom to turn to. Responsibility shifts between stakeholders – developers, product managers, legal counsel, and third-party service providers – without a stable point of accountability. This ambiguity complicates efforts to embed privacy into the development process and impedes the establishment of effective compliance mechanisms [31].

- **Developers have difficulty understanding privacy requirements.** Developers often face significant challenges in understanding and interpreting privacy requirements. A key obstacle to GDPR onboarding is the general lack of familiarity with its principles among developers [1]. Rather than engaging directly with privacy concepts, developers frequently adopt the vocabulary and mindset of data security to approach privacy-related tasks, which can lead to misaligned implementations [10]. In practice, many developers lack a comprehensive understanding of the behavior and implications of third-party SDKs they integrate into their applications [2]. Core principles such as data minimization, fairness, and data protection by design and by default are perceived as abstract and difficult to operationalize in everyday development practices. This lack of clarity is especially problematic as developers have legal responsibilities under data protection law if they become a data controller, see Section 4.2.2 and [8].

Relationships created by the software development life cycle and the state of the software development market. Developers have implicit relationships with various third-party software providers since today's developers (here, as code users) depend on third-party services and in most contexts, it has become almost impossible to build software without third-party components. However, developers also rely on third-party privacy and compliance solutions that are not always effective. Finally, third-party services are shipped with privacy-unfriendly default settings and sometimes manipulate developers towards non-compliance.

- **Developers depend on and trust third-party services.**
  Web ecosystem: Websites trust and heavily depend on various types of third-party tools and services. Research surveying top-500 popular websites in 50 countries together covering all inhabited continents found out that dependencies on a third-party Domain Name System (DNS) and on Content Distribution Networks (CDNs) or Certificate Authorities (CAs) provider vary widely around the world, ranging from 19% to as much as 76% of websites, across all countries [16]. However, there is a highly concentrated market of third-party providers: three providers across all countries serve an average of 92% of websites and Google, by itself, serves an average of 70% of the surveyed websites. Websites rely on third parties for useful and visible content: Ikram et al. found that 50% of first-party websites render content that they did not directly load [14]. Additionally, multiple studies have measured the presence of third-party tracking on thousands of websites even if it is not always clear whether such trackers were included directly by the website developers or included in functional third-party content. Moreover, new forms of such tracking is being included: recent research [20, 39] has shown that "nearly 90% of all websites use at least one tracking first-party cookie, 96% of which are in fact set by third-party scripts running in a first-party context" [38]. Recent work shows that even if website owners include very popular third-party services, the services often provide privacy policies and technical documentation that do not match each other, and moreover do not match the actual data that will be collected from end users by these third-party services. This was studied in the context of Google Tag Manager [18].

Mobile app developers: Mobile app stores require app developers to include a privacy policy. Yet, they do not provide information about how to write a privacy policy. Developers therefore often need to comply with regulations and requirements without knowing what needs to go into the privacy policies [34]. There is minimal research on helping developers craft privacy policies. The lack of support can be seen in the wild, where there are still numerous apps without privacy policies as well as privacy policies that contain misleading and contradictory statements [35]. Ad networks do not always provide guidance about what a developer should include in a privacy policy [33].

- **Developers delegate privacy and compliance to third-party solutions that are often not effective**.

  Compliance solutions: Website owners use "GDPR compliance solutions" that would scan their websites and identify privacy violations, such as presence of trackers, believing that these solutions would ensure compliance. In practice, such compliance solutions contain both false positives – deceiving website publishers into believing that a consent banner is needed on an empty website without trackers [36] – and false negatives – compliance solutions only scan cookies, but miss other Web tracking technologies, such as browser fingerprinting and hence data processed without legal basis [25]. In addition, only few tools help developers configure privacy solutions and disclosures, like user-friendly consent popups and accurate privacy policies, and even fewer tools take into account both regulations and the current behaviours of common third-party APIs [30].

  Privacy solutions: Research shows that developers believe that privacy-oriented libraries would provide an effective privacy solution without truly understanding functionality of such libraries. Song et al. [28] used qualitative methods and mental models approaches to analyze the differences between conceptual models used to design open-source Differential privacy (DP) libraries and mental models of DP held by users. They found that comparing developers' conceptual models with users' mental models elucidates crucial gaps between theory and practice of DP libraries.

- **Rather than privacy by default, there is manipulation against privacy settings and towards non-compliance.**

  No privacy by default in third-party libraries: Developers tend to follow the guidelines and requirements provided by the platforms [27, 35]. However, online services are often built with no compliance features embedded. Third-party services (libraries, SDKs) that Web/app developers rely on are very often not privacy-preserving and not privacy compliant by default [23]. Moreover, developers are reluctant to change the default settings: for example, in the context of mobile app SDKs, developers largely keep ad networks' SDK default configuration [19].

  Manipulation of developers by third-party tools: Developers (here, those using the code) are shown to be manipulated by the user interfaces of the compliance tools, such as Consent Management Platforms (CMPs) provided by third-parties, who nudge developers towards installing non-compliant cookie consent banners [36].

Expertise and developers' relationship to it. The following challenges regard the impact of the relationships in which developers seek and use advice.

- **Communication gaps between developers and privacy experts abound.** Communication gaps between developers and (legal) privacy experts (arising already from basics such as mismatches between legal and technological terminology and conceptual systems [26]) hinder the effective implementation of current GDPR compliant software [12]. In addition, developers tend to omit contacting privacy experts for support to

avoid burdening them [13]. The creation of knowledge resources such as repositories is often viewed as a remedy. However, such repositories, which – in spite of creating initial enthusiasm on all sides – tend to turn out costly and laborious to maintain, and thus tend to give rise to projects that are discontinued [26].

- **Developers use questionable online sources for code/third-party components and legal information.** As noted above, developers have difficulty understanding privacy requirements. They often turn to online sources, both for legal information and for code meant to implement legal requirements, in order to be compliant with data protection law. Traditionally, developers have turned to communities and their sites such as StackOverflow, and also Google or YouTube, for guidance [31]. These platforms have become informal spaces where developers discuss privacy, ethics, and values, highlighting a gap in accessible, authoritative resources [9].
  Currently, people are increasingly turning to chatbots for advice. A steep decrease in StackOverflow usage was noted already in 2023 [5]; this trend continues [21]. As a consequence, community knowledge becomes centralized and subject to non-transparent processing. More specifically, question and advice texts are processed as (re)training data of privately owned and difficult-to-scrutinize LLMs. In addition, by virtue of the personalized and ephemeral nature of chatbot dialogues, this knowledge becomes inaccessible for oversight by experts.
  Various research has examined how the release of ChatGPT-3.5 has affected user engagement on StackOverflow and other StackExchange platforms. A recent empirical study [11] arrived at a cautiously optimistic conclusion about interactions on community platforms and the likely quality of advice: While overall activity on StackOverflow declined sharply, dropping to less than 30% of its pre-ChatGPT levels, the effects varied across different types of users. Low-reputation users, who often ask basic questions, significantly reduced their participation, likely turning to Generative AI (GenAI) tools such as ChatGPT for quick answers. In contrast, high-reputation users maintained or slightly shifted their activity, engaging more with each other and focusing on more advanced, peer-level discussions. This suggests that GenAI may play a dual role: automating basic Q&A for novices while complementing expert-level knowledge sharing, which may help sustain valuable online communities and high-quality training data for AI.
  The quality of advice given by chatbots may be affected by various factors. For example, provider bias has been observed: GenAI models show systematic preferences for services from specific providers in their recommendations (e.g., favoring Google Cloud over Microsoft Azure) [40].
- **Further research challenges arise on a meta-level.** In addition to the problems for products and processes that are created by developers drawing on questionable advice, this last point also presents a methodological research challenge. In a world where developers are using GenAI tools instead of Stack Overflow for advice seeking, it is unclear how researchers can observe and measure their behavior. In addition, it is getting even more difficult to research artifacts/objects they are creating as was done for StackOverflow because these resources are not accessible any more and – due to different contexts – not reproducible independently. Also, it is not clear any longer which resources were produced by humans and which by GenAI. Finally, chatbot interactions allow for richer advice (tailored to personal pre-knowledge and scenario at hand) than generic advice, so it is unclear what the benchmarks are for measuring "quality": is it "correct or wrong answers" or "efficiency of interaction" or "specificity of advice"? Which metrics are appropriate and which ones can be measured?

- **Who should provide guidance for which types of developers?** Currently, there are several actors that provide privacy guidance for developers, such as regulatory bodies who provide recommendations and guidelines for best practices to implement the data protection law. In addition, companies, and standardization bodies (such as NIST) also provide guidance.

4.2.4 Key Research Questions

1. Responsibility: Who should be responsible for privacy compliance according to their role?
   - How to identify different user groups we consider "developers" in a specific context?
   - Should third-party providers be responsible since they both control the tools and data collection, even if they are not the ones controlling the primary application and therefore may argue that they are not responsible?
   - What organizational or cultural factors influence how developers understand and engage with privacy requirements – or don't?
2. Knowledge: What knowledge gaps and misconceptions do developers have about privacy, and how do these shape their practices?
   - What kinds of legal information do developers need but often lack?
   - What are the common misunderstandings or contradictions developers hold about privacy and user expectations?
   - How do developers' assumptions about user attitudes influence their privacy-related decisions?
3. Delivery of Guidance: What forms of guidance and support mechanisms are most effective in helping developers implement privacy practices?
   - What forms of guidance (lightweight or embedded) are most useful in practice?
   - What prevents developers from consulting privacy experts or using available resources?
   - Is it necessary to build closer relationships between developers and privacy experts?
   - Can AI tools meaningfully support developers in privacy tasks, and how reliable are they?
   - What form of guidance from third-party services would be useful to help developers understand the privacy implications in their decision-making?
4. Evaluation: How can we evaluate the effectiveness of privacy guidance and support for developers?
   - What criteria or metrics should be used to measure success (e.g., compliance, usability, adoption)?
   - What kinds of experimental or observational studies can assess whether developers actually use and benefit from guidance?
   - How can we simulate or test real-world developer behavior in privacy-critical scenarios?
5. Meta-analysis: In an age where advice seeking is happening in private spaces (chatbots), what are the appropriate ways to study developer behavior and traces they left in the public space (e.g., artifacts/objects – benchmark on correct or wrong)?
   - How to research what chatbot interaction looks like?
   - Can we expect developers to donate their data (e.g., GenAI-prompting history)? We need to consider research ethics in this context.
   - How about intellectual property issues and authorship?

4.2.5 Solution ideas and directions

Solutions range from foundational education to increasingly in-the-moment support.

1. Build capacity and educate programmers with a privacy-aware mindset:
   - The next generation of computer scientists needs holistic education and effective training, in contrast to generic annual privacy training. These efforts should hone a certain mindset rather than focus on teaching yet another modular skill.
   - This education and training faces uncertainty around the value of skill sets amid the increasing use of generative AI. A large-scale deskilling regarding critical thinking [17] may also endanger democratic agency. The risk of deskilling should be kept in mind, researched, and potentially countered.
2. Equip developers with safe software engineering practices:
   - Support mechanisms should deliver clarity of privacy goals through requirements engineering and its translation into software architecture, considering the mental models that developers have of end-users [7].
3. Provide contextual support:
   - IDEs should offer easily accessible contextual programming aids for implementing privacy – potentially leveraging AI-based code-generation and AI-based advice (akin to secure development lifecycle and existing tools like Copilot). Barriers to use need to be lowered, and quality assurance of the advice given to developers strengthened – regardless of how and where in the messy software development process people seek advice.
   - Part of the solution might be a repository of privacy-information-seeking behavior of developers that can be studied by researchers (especially important in a present and future in which developers seek advice from chatbots, i.e., not in the open, and every developer potentially getting personalized different kinds of advice).
   - Good privacy must be the easy choice: Dark patterns must not be perpetuated and there should be privacy-friendly options when consuming and updating third-party code [32], such as coarse-grained defaults and code snippets for location APIs [15]. Limited monetisation options can be a struggle towards better privacy [6].
   - Higher transparency regarding data practices (collection, processing, storage etc.) in the market of software products and for third-party libraries is needed, so that developers are empowered to choose wisely which libraries or services they link into their software. Appropriate metrics can support end-to-end accountability aided by transparency along the software supply chain. Similar consumer-oriented portals already exist, such as Mozilla's "Privacy not included" (which is a "naming and shaming portal for IoT devices").

4.2.6 Roadmap Development

Directions for for academia, policy-makers, industry should proceed along the following questions.

1. What immediate research questions need to be answered?
   - What motivates developers to care for privacy in different contexts?
   - How can we integrate privacy support in developers's day-to-day work?
   - What kind of guidance is effective for different kinds of developers?

- What kinds of privacy-related development tasks/requirements can be supported with existing off-the-shelf GenAI tools or coding aids that are used today? Where do they make mistakes?
- What could motivate third-party service providers to provide services with privacy-preserving defaults?

2. What are challenges and questions to be answered in the next three years?
   - How can GenAI tools be adapted/improved to provide helpful guidance for developers?
   - What role does the current and foreseeable market power of a small number of commercial Big GenAI tools play in the answer to the previous question?
   - How to motivate developers in public bodies (administration/governmental institutions) to care for privacy?
   - How to make privacy a priority task for developers?
   - How to motivate third-party service providers to provide compliant services?
3. What are challenges and questions to be answered within the next decade?
   - What are the effects of the quickly evolving GenAI tools, which changes how developers work and who develops software?
4. What skills and collaborations are necessary to address them?
   - Collaboration between scholars from computer science and law, psychology, ethics, economics
   - Collaboration with practitioners (controllers, processors) and data protection authorities
   - Collaboration with legal experts that consult companies on data protection requirements
   - Collaboration with industry bodies (like VDI/VDE or IHK/Handwerkskammer in DE) that advise organizations
5. What tools, infrastructure, funding, or incentives are needed?
   - Tools that assist developers in making more privacy-friendly decisions
   - Actual testbeds to try some new interventions. (technical support and funding for support staff)
6. What roadblocks may slow down or discourage research progress?
   - Perceived poor enforcement of GDPR violations may disincentivize developers to care for privacy.
   - Changes to GDPR that increase or reduce requirements for controllers.

Methodologically, this may work best by working with case studies. Candidate areas for developers are consent management products and tag managers.

References

1 Abdulrahman Alhazmi and Nalin Asanka Gamagedara Arachchilage. I'm all ears! Listening to software developers on putting gdpr principles into software development practice. *Personal and Ubiquitous Computing*, 25(5):879–892, 2021.

2 Noura Alomar and Serge Egelman. Developers say the darnedest things: Privacy compliance processes followed by developers of child-directed apps. *Proceedings on Privacy Enhancing Technologies*, 2022.

3 Rebecca Balebako and Lorrie Cranor. Improving app privacy: Nudging app developers to protect user privacy. *IEEE Security & Privacy*, 12(4):55–58, 2014.

4 Partha Das Chowdhury, Joseph Hallett, Nikhil Patnaik, Mohammad Tahaei, and Awais Rashid. Developers are neither enemies nor users: they are collaborators. In *2021 IEEE Secure Development Conference (SecDev)*, pages 47–55. IEEE, 2021.

5 R Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS nexus*, 3(9):pgae400, 2024.

6 Anirudh Ekambaranathan, Jun Zhao, and Max Van Kleek. "Money makes the world go around": Identifying barriers to better privacy in children's apps from developers' perspectives. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 46:1–46:15. ACM, 2021.

7 Anirudh Ekambaranathan, Jun Zhao, and Max Van Kleek. How can we design privacy-friendly apps for children? Using a research through design process to understand developers' needs and challenges. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29, 2023.

8 Michèle Finck. Cobwebs of control: the two imaginations of the data controller in EU law. *International Data Privacy Law*, 11(4):333–347, 2021.

9 Rohan Grover. Encoding privacy: Sociotechnical dynamics of data protection compliance work. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2024.

10 Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*, 23(1):259–289, 2018.

11 Babak Heydari and Negin Maddah. The shifting dynamics of online knowledge platforms and the implications for generative ai sustainability, 2025. Available at SSRN: `https://ssrn.com/abstract=5117087`.

12 Stefan Albert Horstmann, Samuel Domiks, Marco Gutfleisch, Mindy Tran, Yasemin Acar, Veelasha Moonsamy, and Alena Naiakshina. "Those things are written by lawyers, and programmers are reading that." Mapping the communication gap between software developers and privacy experts. *Proceedings on Privacy Enhancing Technologies*, 2024.

13 Stefan Albert Horstmann, Sandy Hong, David Klein, Raphael Serafini, Martin Degeling, Martin Johns, Veelasha Moonsamy, and Alena Naiakshina. "Sorry for bugging you so much." Exploring developers' behavior towards privacy-compliant implementation. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 1215–1233. IEEE, 2025.

14 Muhammad Ikram, Rahat Masood, Gareth Tyson, Mohamed Ali Kaafar, Noha Loizon, and Roya Ensafi. The chain of implicit trust: An analysis of the web third-party resources loading. In *The World Wide Web Conference*, pages 2851–2857, 2019.

15 Shubham Jain, Janne Lindqvist, et al. Should I protect you? Understanding developers' behavior to privacy-preserving APIs. In *Workshop on Usable Security (USEC'14)*, 2014.

16 Rashna Kumar, Sana Asif, Elise Lee, and Fabian E Bustamante. Each at its own pace: Third-party dependency and centralization around the world. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–29, 2023.

17 Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–22, 2025.

18 Gilles Mertens, Nataliia Bielova, Vincent Roca, and Cristiana Santos. You can't trust your tag neither: Privacy leaks and potential legal violations within the google tag manager. In *EuroS&P 2025-10th IEEE European Symposium on Security and Privacy*, 2025.

19 Abraham H Mhaidli, Yixin Zou, and Florian Schaub. "We can't live without them!" App developers' adoption of ad networks and their considerations of consumer risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 225–244, 2019.

20 Shaoor Munir, Sandra Siby, Umar Iqbal, Steven Englehardt, Zubair Shafiq, and Carmela Troncoso. Cookiegraph: Understanding and detecting first-party tracking cookies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3490–3504, 2023.

21 Seokran Park and Dongyeon Kim. Impacts of AI-based search on user engagement: Evidence from stack overflow's overflowai. In Michael D. Myers, Rose Alinda Alias, Wai Fong Boh, Robert M. Davisin, Barney Tan, and Nor Zairah Ab Rahim, editors, *29th Pacific Asia Conference on Information Systems, PACIS 2025, Kuala Lumpur, Malaysia, July 6-9, 2025*, 2025.

22 Article 29 Data Protection Working Party. Guidelines on transparency under regulation 2016/679, 17/en wp260 rev.01, April 2018.

23 David Rodriguez, Joseph A Calandrino, Jose M Del Alamo, and Norman Sadeh. Privacy settings of third-party libraries in Android apps: A study of Facebook SDKs. *Proceedings on Privacy Enhancing Technologies*, 2025.

24 Arianna Rossi, Rossana Ducato, Helena Haapio, and Stefania Passera. When design met law: Design patterns for information transparency. *Droit de la Consommation = Consumenterecht: DCCR*, (122–123):79–121, 2019.

25 Cristiana Santos, Midas Nouwens, Michael Toth, Nataliia Bielova, and Vincent Roca. Consent management platforms under the GDPR: processors and/or controllers? In *Annual Privacy Forum*, pages 47–69. Springer, 2021.

26 Stefan Schiffner, Bettina Berendt, Triin Siil, Martin Degeling, Robert Riemann, Florian Schaub, Kim Wuyts, Massimo Attoresi, Seda Gürses, Achim Klabunde, et al. Towards a roadmap for privacy technologies and the general data protection regulation: a transatlantic initiative. In *Annual privacy forum*, pages 24–42. Springer, 2018.

27 Katie Shilton and Daniel Greene. Linking platforms, practices, and developer ethics: Levers for privacy discourse in mobile application development. *Journal of Business Ethics*, 155(1):131–146, 2019.

28 Patrick Song, Jayshree Sarathy, Michael Shoemate, and Salil Vadhan. "I inherently just trust that it works"': Investigating mental models of open-source libraries for differential privacy. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–39, 2024.

29 Alina Stöver, Nina Gerber, Henning Pridöhl, Max Maass, Sebastian Bretthauer, Matthias Hollick, Dominik Herrmann, et al. How website owners face privacy issues: Thematic analysis of responses from a covert notification study reveals diverse circumstances and challenges. *Proceedings on Privacy Enhancing Technologies*, 2023.

30 Ruoxi Sun and Minhui Xue. Quality assessment of online automated privacy policy generators: an empirical study. In *Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering*, pages 270–275, 2020.

31 Mohammad Tahaei, Marvin Ramokapane, Tianshi Li, Jason I Hong, and Awais Rashid. Charting app developers' journey through privacy regulation features in ad networks. In *The 22nd Privacy Enhancing Technologies Symposium*, pages 33–56. De Gruyter Open Ltd., 2022.

32 Mohammad Tahaei and Kami Vaniea. A survey on developer-centred security. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 129–138. IEEE, 2019.

33 Mohammad Tahaei and Kami Vaniea. "Developers are responsible": What ad networks tell developers about privacy. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2021.

34 Mohammad Tahaei, Kami Vaniea, and Awais Rashid. Embedding privacy into design through software developers: Challenges and solutions. *IEEE Security & Privacy*, 21(1):49–57, 2022.

35 Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. Understanding privacy-related questions on stack overflow. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

36 Michael Toth, Nataliia Bielova, and Vincent Roca. On dark patterns and manipulation of website publishers by CMPs. *Proceedings on Privacy Enhancing Technologies*, 2022(3):478–497, 2022.

37 Christine Utz, Sabrina Amft, Martin Degeling, Thorsten Holz, Sascha Fahl, and Florian Schaub. Privacy rarely considered: Exploring considerations in the adoption of third-party services by websites. *Proceedings on Privacy Enhancing Technologies*, 2023.

38 Yash Vekaria, Yohan Beugin, Shaoor Munir, Gunes Acar, Nataliia Bielova, Steven Englehardt, Umar Iqbal, Alexandros Kapravelos, Pierre Laperdrix, Nick Nikiforakis, Jason Polakis, Franziska Roesner, Zubair Shafiq, and Sebastian Zimmeck. SoK: Advances and open problems in web tracking. *CoRR*, abs/2506.14057, 2025.

39 Yash Vekaria, Benjamin Standaert, Max Ostapenko, Abdul Haddi Amjad, Yana Dimova, Shaoor Munir, Chris Böttger, and Umar Iqbal. Chapter 10: Privacy. In *The 2024 Web Almanac*. HTTP Archive, 2024. [Online]. Available: `https://almanac.httparchive.org/en/2024/privacy`.

40 Xiaoyu Zhang, Juan Zhai, Shiqing Ma, Qingshuang Bao, Weipeng Jiang, Qian Wang, Chao Shen, and Yang Liu. The invisible hand: Unveiling provider bias in large language models for code generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21376–21403, Vienna, Austria, July 2025. Association for Computational Linguistics.

4.3 AI for Privacy, Privacy for AI

Benjamin Berens (Karlsruhe Institute of Technology, DE, benjamin.berens@kit.edu)
Simone Fischer-Hübner (Karlstad University, Chalmers University of Technology & University of Gothenburg, SE, simone.fischer-Hübner@kau.se)
Andreas Gutmann (University College London, UK)
Bailey Kacsmar (University of Alberta – Edmonton, CA, kacsmar@ualberta.ca)
Agnieszka Kitkowska (Jönköping University, SE, agnieszka.kitkowska@ju.se)
Marc Langheinrich (Università della Svizzera italiana – Lugano, CH, marc.langheinrich@usi.ch)
Mainack Mondal (Indian Institute of Technology Kharagpur, IN, mainack@cse.iitkgp.ac.in)
Elissa Redmiles (Georgetown University – Washington DC, US, elissa.redmiles@georgetown.edu)

4.3.1 Introduction

The current human-AI interaction paradigm is making computational machines more pervasively trained and deployed on much more data (than previous generations of AI). This raises both the opportunity to use AI tools to enhance societal outcomes, including privacy, and threats from the exacerbation and introduction of risks and harms. We posit that going

forward, we must address the question "how does human-centered privacy change in a world where AI-based automation (e.g., AI-agents) is incorporated into the many facets of digital life, for better or worse?"

Risks and harms may come as a result of decisions made by individuals (who use an AI-employing system), by others (collateral privacy through inferences made about you even as a non-user), and by society (to prioritize an AI capability over a threat to some individuals). Therefore, to understand these different sources of harm, we first need to surmise if existing privacy and behavioral decision-making theories, and decision-making consequences, explain and predict such decisions as well as potential privacy issues that consequently emerge. If not, we need to update these theories and frameworks to better explain and predict potential privacy issues that can emerge from individual human-AI interaction and societal decisions about AI deployment. From there we can develop a better understanding of decision making, risks and threats specific to this domain, as well as how AI can be used to benefit society helping rather than hindering privacy in our digital spaces.

Finally, we emphasize that while AI poses new challenges in the space of human-centered privacy, it also offers positive potential for providing usable end-user guidance or even for automated or semi-automated privacy decisions.

4.3.2 Problems/Challenges Description

Transferability of Existing Theories and Methods. The existing models of decision-making and behavioral outcomes in the context of privacy assume that the same factors that enable individual information disclosures for more traditional technologies also apply to the context of AI-based technologies, e.g., when individuals interact with genAI [2] embedded in chatbots or social robots. However, we have little data or theory investigating or explaining how AI-based technologies might actually be influencing the behaviors of individuals or groups, as well as how such influences can change over time and across different contexts (e.g., usage scenarios).

Creating interaction-grounded frameworks that could provide us with an opportunity to better understand how individual and societal privacy decisions and behaviors change because of AI used in the systems requires handling additional challenges, as today AI transforms privacy decision-making by automating decisions that influence human choices through design, personalization, or implementation of anthropomorphic features. Overall, these features have the potential to influence the trust users put in AI-driven systems. It remains unknown whether the effects of such influences on privacy decisions and behaviors are positive or, on the contrary, negatively affect user privacy and control over their data.

Threats, Risks, and Harms. A fundamental question faces privacy for AI: what threats are essentially exacerbations of prior known privacy concerns, through AI's mechanism of scale, and what threats are novel or considerably altered as a result of capabilities and scalability introduced by human-AI interaction. Prior work on privacy of AI has centred heavily on model-focused attacks on data confidentiality, integrity, correctness, memorization [17], extraction [15], and inference attacks [16], leaving less explored the impact of human-AI interaction and decision-making. Among the threats that exist, or may come to be, we need to assess the risk (impact $\times$ likelihood) each poses, build robust threat models that map the harms that may materialise from these risks (considering differences in individuals and circumstances), and design sociotechnical mitigations into products / processes / society (laws, regulators, etc).

[2] genAI throughout the rest of the chapter refers to generative AI models.

AI exacerbates existing privacy risks. First, AI anthropomorphism may exacerbate people's greater willingness to share information with machines. Second, AI exacerbates existing surveillance risk, including collateral privacy – the impact of others' data/interactions on an individual's risk. Third, AI exacerbates data risks of both extraction (intentional) and leakage (intentional or unintentional). Examples include leaks through information sharing during benign use of benign tools and through interaction with malicious tools.

AI also presents at least two novel privacy risks. First, agentic AI (action-taking agents) may be used maliciously to execute and scale attacks, e.g., social engineering, or may leak private information during benign action. Second, AI capabilities allow for increased plausibility in the generation of recognizable digital replicas (i.e., deepfakes).

A cross-cutting theme is that both existing and novel risks will be amplified by a desire for greater utility: more capabilities, with higher quality. Given the probabilistic nature of AI, some threats cannot be defended against in the near term. Thus, mitigation may include not only technical solutions but societal-level trade-offs between tool capabilities and corresponding indefensible risks and/or identification of social-norm shaping interventions to bound the use of capabilities.

Using AI for usable personalized privacy. In recent years, different AI tools, including personalized privacy assistants, have been developed for assisting and guiding users with privacy decision making (e.g., [8]). (Semi-)automation of privacy decisions founded on AI-based personalised guidance can enhance usable privacy management for users who are often overwhelmed with privacy decision requests, by proposing or enforcing privacy decisions meeting these users personal preferences. AI tools offering personalised privacy assistance can also consider different preferences that users with different demographic backgrounds (e.g. culture, gender, age) typically have, and can enhance accessibility by adapting privacy information to specific user needs.

On the other hand, automation of privacy decisions as well as "privacy nudges" also raise ethical and legal questions, especially regarding the users' autonomy and control over their data, which is an essential privacy principle highlighted in Recital 7 GDPR [3]. Some privacy decisions, such as consent, legally require human actions, i.e., cannot be fully automated. Another problem is that users may overly trust AI-proposed decisions (automation bias). While AI-based tools have been developed that propose decisions matching users' expectations with a very high accuracy, many AI techniques are probabilistic and hence cannot provide any guarantees that the proposed actions reflect the users' wishes. In addition, LLM-based approaches may provide information that is inaccurate or plain wrong. Hence, finding the right balance between AI-based personalised decision guidance and enforcement, and user involvement – keeping users in the loop – constitutes a challenge.

Moreover, ML-based personalized privacy assistants typically analyze the users' attitudinal or behavioral privacy preferences, metadata or content, or other data types. Hence, they in turn need to process personal data and user profiles. Therefore, it is essential that personalised privacy assistants themselves are developed following a privacy-by-design approach [28], especially if the models are not trained locally on the users' devices and under the users' control. Most research work on personalised privacy assistants does not (or not adequately) address such crucial privacy-by-design aspects.

AI-based tools can also be used for personalizing how privacy policies are communicated (e.g., format, language, style). This can improve accessibility for diverse users, including vulnerable groups. However, simplified or voice-based communication may not meet GDPR

[3] https://gdpr-info.eu/recitals/no-7/

requirements (Articles 12, 13) [29]. And if compromised, AI-based privacy assistants could manipulate users into disclosing sensitive information or disclose more information than they would have disclosed otherwise, which constitutes a deceptive design (dark pattern) that is illegal in some jurisdictions (e.g., in the EU Digital Services Act, which is reflected also in AI Act, Digital Markets Act, and Consumer Rights Directive).

4.3.3 State of the Art

Different fields (HCI, AI, SWE, law, and more) have each developed disparate definitions of safety as well as classification taxonomies and evaluation frameworks for privacy risks and harms. Some of these build upon prior theories of privacy, such as using Solove's framework [18], while others introduce their own structure. However, while these provide organization to the domain, they each use different conceptual terminology, different categorisations, and largely remain at an overview level, aiming to capture the overall field. This creates challenges and limitations in terms of applicability for those creating, deploying, or overseeing these systems as: (i) By trying to cover the breadth of the space, nuance cannot be captured, (ii) these frameworks are often rooted in academic/theoretical/philosophical perspectives such that those wanting to apply these insights need to translate and transfer them to match their requirements. Unfortunately, these can often not transfer well or easily to such industry processes and priorities (e.g. in regards to when or how harms must be mitigated).

We know from usability foundations that adoption rises when "things" are designed with end-user requirements and use cases in mind from the start, not added at the end. Regulators and governments (for instance EU enacted, Canada tabled, UK's AI Bill tabled) have similarly tried to capture risk as either:

- All AI (which has contentious definitional bounds)
- AI for specific tasks, with some using domain-specific or industry-specific existing guidelines or laws as baselines and others trying to create new ones that specifically target "AI in X"

While we can articulate notions of privacy risk conceptually in this space, which concrete harms will follow from them is less clear, difficult to measure with existing tooling and models, which correspondingly leads to contention as to next steps for all stakeholders and relevant parties.

Following the human-centric approach, to ensure a comprehensive understanding of both the positive and negative effects of AI on people, one must first gain a thorough understanding of how individuals and groups make decisions regarding their privacy and security when using technologies that incorporate AI. Traditional psychological models like the Theory of Reasoned Action (TRA) and Theory of Planned Behavior (TPB) fall short in explaining how people make privacy decisions under uncertainty or cognitive overload [19, 20, 21, 22]. Dual-process theories – System 1 (S1) and System 2 (S2) – provide a more comprehensive framework [22, 21]. S1 is fast, intuitive, and driven by heuristics (e.g., affect heuristic), while S2 is slow, analytical, and effortful. Most everyday decisions rely on S1, which is sensitive to various peripheral cues (e.g., visuals), heuristics and biases, and other factors that might affect cognitive processing of information (e.g., time pressure). Privacy decision-making models build on psychological theories, such as TPB, TRA, or dual process theories, at times combining them into more holistic but theoretical models. One such model, APCO (Antecedents–Privacy Concerns–Outcomes), centers privacy concerns while acknowledging background factors like personality, culture, trust, and skills [23]. This conceptual macro-model incorporates System 1 thinking, recognizing the role of fast, intuitive judgments in privacy choices. Following the model, privacy concerns seem to play a central role to

decision-making, yet, they might be hard to assess and various scales to measure them were developed (e.g., Internet Users's Information Privacy Concerns (IUIPC) [24], Mobile Users Privacy Concerns [25], and even Social Robots Privacy Concerns [26]). AI development at a fast pace resulted in the first attempts to measure concerns in this context with the Privacy Concerns related to AI Misuse (PC-AIM) scale [27], which builds on existing models like the IUIPC but adds dimensions like data permanence, profiling, reduced judgment, and algorithmic bias. However, more studies are needed to validate its accuracy and applicability to various contexts of interactions with AI-based technologies. Considering the complexity of decision-making processes, other measurement instruments/methods for latent factors affecting decisions and behavior in the context of privacy and AI might be needed.

In AI interactions, anthropomorphization can influence privacy decisions, sometimes causing discomfort (the "uncanny valley" effect). However, this can be offset by empathetic responses and immediate implicit feedback (e.g., through multi-turn interactions), which increase user comfort and data disclosure. These effects, though impactful, remain under explored, raising concerns about potential privacy and security risks. AI agents often trigger affective heuristics through emotional cues like personalization and human-like behavior, which can override rational risk assessment. As a result, users may underestimate privacy risks and overshare data, especially under cognitive load or time constraints.

Studies show users are often more willing to share sensitive data with AI than with humans – particularly if AI is perceived as non-judgmental [30, 31]. However, if AI seems too powerful or invasive, privacy concerns intensify. Cultural factors also shape responses: U.S. users tend to trust AI more unless data sensitivity is high, whereas Chinese users are more [32]. It is possible that older adults may accept AI, similarly to how they accept social robots [33], if its utility outweighs privacy risks, but still demand control and transparency; however, research is needed to confirm such assumptions.

Regarding the beneficial use of AI in the context of privacy, i.e., using AI to support stakeholders with privacy-related decision-making, a rapid review of the existing literature sees three main research areas emerging:

(A) Using AI to assess privacy issues in to-be-shared social media content (e.g., [2, 3, 4, 5]): The key challenges are the lack of real-world deployments, the issue of cold-start, and the social complexity inherent in such sharing decisions. Multi-party privacy conflicts are also difficult to solve, with or without AI.

(B) Using GenAI/LLMs for privacy policy understanding (e.g., [34, 7] [Tongue and Caragea 2020]): The key challenges here are unreliability of LLMs, the legal validity of consenting to AI-summarized policies, and, of course, the inherent complexity of policies written not for humans but lawyers

(C) AI-based privacy management (e.g., decision making) (e.g., [10, 6][Bhattacharya, 2024]).

Area (A) has traditionally seen much work due to the importance of social media. Since 2023, area (B) is growing quickly. Area (C) is just emerging.

4.3.4 Key Research Questions

We identified four key research questions in this domain. This report focuses on the first three. However, we include the fourth as we also need to communicate amongst stakeholders all the aspects of the other three research questions.

- RQ1: How can we use AI to enhance usable privacy?
- RQ2: How does AI affect (positively/negatively) human privacy decision-making and behavior?

- RQ3: How does AI pose novel or exacerbate existing risks to people's privacy?
- RQ4: How can developers, companies, researchers, or governments (etc.) communicate privacy (cf RQ1-3), both reactively and proactively, (e.g., risks, potential manipulation) in the context of ubiquitous AI-based systems?

4.3.5 Solution ideas and directions

Guided by our research questions and the literature, we synthesize the path forward as Table 1.

4.3.6 Resources required

Interdisciplinary collaboration: the role of social and psychological factors necessitates collaborations with at minimum Psychology and Sociology. Moreover, cooperation with legal experts is needed for eliciting and enforcing legal requirements.

Hardware (GPU, robotics infra, etc.). Assessments may require direct system interaction Secure funding to support longitudinal research and collection of large datasets (perhaps on an ongoing basis)

Data. Evaluation and benchmarking datasets, including multi-turn interaction data Incident tracking for privacy issues/harms/consequences & regulation for data access to commercial data (cf. research access provisions in EU and UK regulation of online safety)

4.3.7 Roadblocks

Speed of innovation. High volatility of everything with AI (possibly relating to views on "outdatedness" as well as changes in applications and mitigations of issues)

Disconnection For collaborative efforts between technical, legal, social sciences, and design knowledge bases, there are correspondingly interdisciplinary challenges in the collaboration between people with deep technical AI knowledge and other relevant skill sets (law, HCI, psychology, sociology).

References

1 Morel, V., Iwaya, L. & Fischer-Hübner, S. AI-driven Personalized Privacy Assistants: a Systematic Literature Review. *IEEE Access*. (2025)
2 Freiberger, V., Fleig, A. & Buchmann, E. "You don't need a university degree to comprehend data protection this way": LLM-Powered Interactive Privacy Policy Assessment. *Proceedings Of The Extended Abstracts Of The CHI Conference On Human Factors In Computing Systems*. pp. 1-12 (2025)
3 Freiberger, V., Fleig, A. & Buchmann, E. Explainable AI in Usable Privacy and Security: Challenges and Opportunities. *ArXiv Preprint ArXiv:2504.12931*. (2025)
4 Hamid, A., Samidi, H., Finin, T., Pappachan, P. & Yus, R. GenAIPABench: A benchmark for generative AI-based privacy assistants. *ArXiv Preprint ArXiv:2309.05138*. (2023)
5 Aydin, I., Diebel-Fischer, H., Freiberger, V., Möller-Klapperich, J., Buchmann, E., Färber, M., Lauber-Rönsberg, A. & Platow, B. Assessing Privacy Policies with AI: Ethical, Legal, and Technical Challenges. *ArXiv Preprint ArXiv:2410.08381*. (2024)
6 Najana, M., Balakrishnan, A. & Bhattacharya, S. Conceptualizing Copilots as Privacy Assistants: A Theoretical Framework. (2024)
7 Ayci, G., Sensoy, M., özgür, A. & Yolum, P. Uncertainty-aware personal assistant for making personalized privacy decisions. *ACM Transactions On Internet Technology*. **23**, 1-24 (2023)

Table 1 The roadmap of strategies towards advancement in the AI for Privacy and Privacy for AI domain per research question.

Horizon	Strategy	RQ1	RQ2	RQ3
Immediate	• Establish threat models that map risks to harms	X		X
	• Validate existing theoretical and empirical models of human-centered privacy in the context of emerging AI capabilities, e.g., re-evaluate the impact of novel surveillance capabilities, future power dynamics, and advanced interaction dynamics on privacy expectations, decisions, and achievability	X	X	X
	• Standardize methodologies for considering tradeoffs between PETs and other AI considerations (fairness, representation, etc.)			X
	• Develop privacy-by-design guidelines for personalized AI-based privacy assistants (i.e., where the AI assistant's capabilities require significant access to the user's data)	X		
	• Elicit legal, regulatory, and/or end user requirements	X	X	X
3+ years	• Operationalize risk assessment measurement tools (mixed methods, need qualitative and quantitative, would need both model-centric, and human-interaction style ones) for a range of AI capabilities including agentic AI	X		X
	• Human-centered design templates for semi-automated tools that action privacy on user's behalf that address issues such as accuracy, semantic understanding, cold start,–along with providing the corresponding validation mechanism that supports subsequent deployments	X		
	• Clarifying validity of AI-Supported (legal) Consent and Liability	X		
10+ years	• Longitudinal studies of Human-AI interaction & changing modalities (esp. Anthropomorphic effects) on human behavior, and of the effect of AI on general privacy decisions, concerns, and behavior	X	X	X
	• Extending semi-automated tool design templates to address interdependent privacy & multi-user collaboration or conflict	X		
	• Regulation and schemes to enable third-party audits		X	X

8 Zhan, N., Sarkadi, S. & Such, J. Privacy-enhanced personal assistants based on dialogues and case similarity. *European Conference On Artificial Intelligence*. (2023)

9 Ayci, G., özgür, A., Sensoy, M. & Yolum, P. Explain to me: Towards understanding privacy decisions. *Proceedings Of The 2023 International Conference On Autonomous Agents And Multiagent Systems*. pp. 2790-2791 (2023)

10 Morel, V. & Fischer-Hübner, S. Automating privacy decisions-where to draw the line?. *2023 IEEE European Symposium On Security And Privacy Workshops (EuroS&PW)*. pp. 108-116 (2023)

11 Carter, S., D'Aquin, M., Spagnuelo, D., Tiddi, I., Cormican, K. & Felzmann, H. The Privacy-Value-App Relationship and the Value-Centered Privacy Assistant. *ArXiv Preprint ArXiv:2308.05700*. (2023)

12 Ischen, C., Araujo, T., Voorveld, H., Noort, G. & Smit, E. Privacy concerns in chatbot interactions. *Chatbot Research And Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3*. pp. 34-48 (2020)

13 Belen Saglam, R., Nurse, J. & Hodges, D. Privacy concerns in chatbot interactions: When to trust and when to worry. *HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23*. pp. 391-399 (2021)

14 Zhan, X., Carrillo, J., Seymour, W. & Such, J. Malicious LLM-Based Conversational AI Makes Users Reveal Personal Information. *USENIX Security*. (2025)

15 Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. & Others Extracting training data from large language models. *30th USENIX Security Symposium (USENIX Security 21)*. pp. 2633-2650 (2021)

16 Staab, R., Vero, M., Balunović, M. & Vechev, M. Beyond memorization: Violating privacy via inference with large language models. *ArXiv Preprint ArXiv:2310.07298*. (2023)

17 Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S. & Oh, S. Propile: Probing privacy leakage in large language models. *Advances In Neural Information Processing Systems*. **36** pp. 20750-20762 (2023)

18 Solove, D. A taxonomy of privacy. *University Of Pennsylvania Law Review.*, 477-560 (2006)

19 Madden, T., Ellen, P. & Ajzen, I. A Comparison of the Theory of Planned Behavior and the Theory of Reasoned Action. *PSPB*. **18** pp. 3-9 (1992)

20 Ajzen, I. Nature and Operation of Attitudes. *Annual Reviews Of Psychology*. **52** pp. 27-58 (2001)

21 Kahneman, D. A Perspective on Judgment and Choice. *American Psychologist*. **3**, 7-18 (2003)

22 Evans, J. & Stanovich, K. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives On Psychological Science*. **8**, 223-241 (2013)

23 Dinev, T., Mcconnell, A., Smith, H., Dinev, T., Raton, B. & Smith, H. Economics : Thinking Outside the "APCO" Box Systems, Psychology, and Behavioral Economics : Thinking Outside the "APCO" Box. *Information Systems Research*. **26**, 639-655 (2015)

24 Malhotra, N., Kim, S. & Agarwal, J. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*. **15**, 336-355 (2004)

25 Xu, H., Gupta, S., Rosson, M. & Carroll, J. Measuring mobile users' concerns for information privacy. (2012)

26 Jia, S., Chi, O. & Lu, L. Social Robot Privacy Concern (SRPC): Rethinking privacy concerns within the hospitality domain. *International Journal Of Hospitality Management*. **122** (2024,9)

27 Menard, P. & Bott, G. Artificial intelligence misuse and concern for information privacy: New construct validation and future directions. *Information Systems Journal*. **35**, 322-367 (2025)

28 Cavoukian, A. Understanding How to Implement Privacy by Design, One Step at a Time. *IEEE Consumer Electronics Magazine*. **9**, 78-82 (2020,3)

29 Commission, E. Regulation (EU) 2016/679 Of The European Parliament And Of The Council of 27 April 2016. *Official Journal Of The European Union*. (2016)

30 Sohn, S., Labrecque, L., Siemon, D. & Morana, S. Artificial intelligence versus human service agents: How their presence shapes consumer information privacy concerns. *Journal Of Retailing*. (2025)

31 Kim, T., Jiang, L., Duhachek, A., Lee, H. & Garvey, A. Do you mind if I ask you a personal question? How AI service agents alter consumer self-disclosure. *Journal Of Service Research*. **25**, 649-666 (2022)

32 Liu, Y., Yan, W., Hu, B., Lin, Z. & Song, Y. Chatbots or humans? Effects of agent identity and information sensitivity on users' privacy management and behavioral intentions: A comparative experimental study between China and the United States. *International Journal Of Human–Computer Interaction*. **40**, 5632-5647 (2024)

33 Reinhardt, D., Khurana, M. & Acosta, L. "I still need my privacy": Exploring the level of comfort and privacy preferences of German-speaking older adults in the case of mobile assistant robots. *Pervasive And Mobile Computing*. **74** pp. 101397 (2021)

34 Kqiku, L. Privacy-Decision Assisting Techniques. (2024)

4.4 Consent, Control, and Communication

Arianna Rossi (Sant'Anna School of Advanced Studies – Pisa, IT, arianna.rossi@santannapisa.it)
Farzaneh Karegar (Karlstad University, SE, farzaneh.karegar@kau.se)
Florian Alt (LMU München, DE, florian.alt@ifi.lmu.de)
Maija Poikela (Charité – Berlin, DE, maija.poikela@bih-charite.de)
Mark Warner (University College London, UK, mark.warner@ucl.ac.uk)
Sophie Grimme (OFFIS – Oldenburg, DE, sophie.grimme@offis.de)
William Seymour (King's College London, UK, william.seymour@kcl.ac.uk)
Zinaida Benenson (Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), DE, zinaida.benenson@fau.de)

4.4.1 Introduction

We are bombarded with cookie banners which often violate informed consent requirements [1], privacy policies are not written to be understood [2], and real time bidding – the economic basis of the contemporary web – is "structurally difficult to reconcile with European data protection law" [3]. While it is clear that the transparency and consent model for privacy decisions is broken [4], its adoption as the cornerstone of consumer and data protection regulation means that it is here to stay. This calls for the need to envision and design solutions that address the continuing problems of transparency and consent. At the same time, the continued use and development of artificial intelligence has created a data-oriented world where it seems inevitable that data flows are destined to grow. This development can

be beneficial for individuals and societies (e.g., for data-driven medical research), providing new opportunities. At the same time, emerging technologies and novel interaction modalities also pose new threats which the research community must anticipate.

In this context, it is essential that we work together to improve the mechanisms that we have to control how our data is used and how we are asked for consent. In our discussions, we strive to create effective, usable, lawful, and accessible tools for consent, control, and communication around the use of personal data.

We make a distinction between the idea of *consent* as an explicit decision point allowing for positive and negative choices, and *control* as the broader category of mechanisms that people have to influence how their data is used. Notably, whilst consent is always something that the data subject negotiates with a data controller, control as used here encompasses resistance, activism, and any other actions taken by an individual, or some other entity on their behalf, that change how their data is used.

4.4.2 State of the Art

Despite decades of research and regulatory attention, privacy notices and consent forms continue to fall short of their intended purposes: attracting user attention, informing individuals about data practices, enabling comprehension, and ensuring meaningful control over personal data. Although during the years privacy policies of service providers have been improved, they are still often compliance-driven rather than user-centred. For example, privacy policy forms, which are one form of privacy notices, have to fulfil requirements from different stakeholders [5, 6]. While users expect to receive clear, simple information about data practices and privacy controls, service providers employ privacy policies to demonstrate compliance with legal requirements.

These shortcomings are further compounded by usability issues such as cluttered layouts, inconsistent placement of privacy controls, or technical barriers that make opting out difficult [7]. In addition, consent mechanisms on websites and digital platforms often rely on design choices nudging users toward acceptance instead of fostering genuine understanding or control. Empirical studies show that interface manipulations such as default buttons, pre-selected options, or the removal and obscuring of rejection choices can raise acceptance [8]. A (perceived) lack of choice and decoupled notices are other reasons why users do not pay attention to privacy notices [5, 6].

This results in notice fatigue, habituation, post-decision regret, and superficial consent, where users agree without fully understanding the implications [9, 10]. Consequently, although many implementations formally satisfy regulatory requirements, they fail to meet the spirit of informed and freely given consent, thereby undermining the user's ability to exercise meaningful control over personal data.

To partly address the usability issues, researchers have explored ways to improve the usability of privacy notices and consent mechanisms. Proposed solutions include privacy nutrition labels [11], layered and short-form summaries [12, 22], personalised notices [13, 14], icons [15, 16], and even comic-based interfaces [17], among other design patterns [24]. These approaches attempt to make complex data practices more transparent and digestible, but each comes with trade-offs: condensed formats risk oversimplification, while visual representations, such as the meaning behind icons, often require additional user learning. Beyond presentation, other efforts focus on interactive consent, using mechanisms like drag-and-drop, swiping, or checkboxes, to actively engage users with the content, improving both attention and retention [18].

Emerging technologies intensify the challenges in these contexts. Artificial intelligence systems often operate opaquely, making it difficult to communicate how data is processed or how decisions are made, although they have the potential for improving usable personalised privacy and decision-making. In AR/VR and metaverse environments, the traditional notice and choice paradigm is even less effective and sometimes impossible. Immersive systems collect highly sensitive biometric and behavioural data, much of which is invisible to the user. The immersive nature of these platforms also makes interruptions for lengthy consent impractical, raising the risk that consent becomes implicit or illusory.

4.4.3 Case Studies

The working group considered a range of different use cases in order to map out the problem space. Each case study was chosen because it surfaced relevant issues and solutions related to consent, control, and/or communication of data practices.

- **Medical data donation for research purposes.** This included discussion around the values and reasons behind the donation of medical data, the potential positive and negative consequences of doing so, and how this could be effectively, lawfully, and ethically communicated.
- **Consent and control for virtual/mixed reality.** We discussed if and when we should design consent decisions to avoid pulling users out of immersive experiences, as well as alternatives such as making choices before/after immersive moments or through other interaction modalities.
- **Conversational Agents.** We talked about how conversation, particularly when spoken, is a relatively low bandwidth means of interaction, but does present opportunities for more engaging conversations about privacy. It foregrounds a question from multi-layer privacy policies about what should be in the first layer.
- **Data collection by smart cars.** The discussion covered how people don't expect this data collection, leading to cases where consent isn't freely given. There are also tangible consequences that arise from data collected by cars, particularly related to car insurance.

When considering these case studies, two main areas of discussion emerged. The first of these was the lack of communication around consequences for data sharing, making it difficult to answer the question of why people should care about privacy. The second centred on the development of AI agents that (amongst other things) would be tasked with making privacy decisions on our behalf.

4.4.4 Focus I: Consequences of Privacy Decisions

This area of discussion was sparked by an observation in the main session that, as privacy experts, we often struggle to effectively communicate the importance of digital privacy. An obvious response would be to convey the tangible consequences (good and bad) and risks of making a privacy decision, but this is something that we don't currently have a good overview of. To this end, the group considered the specification of a consequences-based and cross-platform privacy decision manager that would help people to make privacy decisions that are aligned with their individual expectations and values. We preliminary call this system 3CM: *Consent, Control and Communication Manager* and imagine it to be similar to existing password managers. This idea is based on discussions of the group "Consequence-based Privacy Decisions" in Dagstuhl Seminar "My Life, Shared" from 2013[4] [19].

[4] https://www.dagstuhl.de/13312

A core component of 3CM would be a database of events and outcomes of sharing data, both positive (e.g., personalisation) and negative (e.g., data breaches). Given appropriate context, an LLM (Large Language Model) would be able to provide explanations of these consequences of sharing or not sharing before a decision is made. These decisions would be stored by the manager for future use and could be applied automatically where directed. This would add functionalities to existing tools such as Consent-O-Matic[5]. The privacy decision manager would learn from the decisions that the user makes, taking into account contextual elements that might impact sharing decisions.

Such a concept is not without challenges. In 2013, one of the main challenges was how the system would present the consequences and learn from past decisions. These challenges remain, but seem to be better manageable with the proliferation of LLMs. There are also implementation difficulties around curating data sources and creating experiences that function across devices. A consequences-based decision manager would not address underlying problems with the notice and consent model, but it would present an opportunity to push back and change the incentives for users (and thus organisations) around the choices they make in relation to their data. Ultimately, it would also need to be *easier to use than not use*, increasing user acceptance. This kind of system also brings with it the question of when or if to ask users if they want to reconsider their decisions, potentially leveraging so-called 'teachable moments' where people have the time and energy to make decisions outside of the task that led to the initial prompt to share data.

4.4.5 Focus II: Agents

Artificial agents are increasingly embedded in our day-to-day activities, whether professional or personal tasks. Especially since the widespread availability of conversational interfaces that enable anyone to interact with Large Language Models (LLMs), such as GPT, it is evident that humans find it useful and beneficial to resort to those instruments, even if their usage is often accompanied by a lack of awareness of the entailed risks. As in other contexts, in this kind of interactions, privacy and security are not the primary task of users. Thus, certain old-standing issues are being re-proposed with emerging technologies, in addition to the new challenges these technologies pose.

The group also discussed possible solutions offered by artificial agents to the so-called "privacy self-management dilemma". A trained artificial agent could take privacy decisions on behalf of the user while it performs other tasks (e.g., refuse profiling cookies while booking a flight). But this scenario raises additional questions, such as whether and to what extent an agent can make decisions on behalf of the user. Even if this could sometimes be determined contextually (e.g., not in high-risk scenarios), still, to make informed decisions, the agent should understand the contextual norms of using personal information. In this regard, it would be similar to having a human assistant. The participants discussed whether it is possible that the agent learns through a process of personalization or if it should be based on the profiles of others, as is already the case for some privacy assistants. Another open question concerns whether a threshold for the acceptable amount of errors that can be made should be set. Such an agent calls into question the very nature of "informed" decisions, a concept which should probably be reformulated. What is interesting is that taking privacy-friendly decisions not only concerns the type and amount of data that is shared or withheld, but also the capacity to make inferences based on such data.

[5] `https://consentomatic.au.dk`

In addition, the group discussed the types of requirements that should be formalized for the development of such an agent. These should concern the user interface (e.g., in terms of timing and modality of communication) and should envision the situations where a decision needs to be deferred to a human. This setting also emphasizes the central role that trust would play: requirements should concern the establishment and maintenance of user trust in the agent and cover cases where a breakdown of trust may occur. As a conclusion, the participants acknowledged the need to also include experts with AI and legal backgrounds to address all these questions meaningfully.

Bringing together the two focuses, the group was keen to explore the creation of a privacy decision agent that collated and explained the potential consequences of different privacy decisions.

4.4.6 Key Research Questions

Based on the discussions during the week, the group compiled a list of key research questions for the research community. These are sorted by topic area and timeline:

4.4.6.1 Human factors issues around consent and control

- Immediate research questions
  1. What are researchers' conceptualisations, and users' mental models of providing, revoking and negotiating consent? How do researchers' and users' views differ?
  2. What is the users' understanding of the current state of privacy policies compared to 10-15 years ago?
  3. Does enhanced user-centred transparency exist, and if so, has it been effective at all? What would more "effective" mean (i.e., noticeable, memorizable, understandable, etc)?
  4. Are users aware of the enhanced control promised in privacy rules and regulations, and to what extent do they exercise their rights? Does "enhanced control" (i.e., more actionable rights that are now available to users) make people feel more in control?
  5. If people feel more in control, do they feel more satisfied with their privacy?
  6. What would a systematic taxonomy of control mechanisms look like (e.g., contacting a DPA; changing your data; right to access, etc)?
- Challenges and questions for the next 3 years
  7. How can informed consent be conceptualised differently from a human factors perspective?
  8. Under which circumstances can the necessity of active decision-making be lifted from a legal, ethical, and human point of view?
  9. Do existing implementations of dynamic consent "work"? For example, do people revise their decisions? If not, why not? Do they feel information fatigue?
  10. How is the right to lodge a complaint implemented across different national DPAs? Who has invoked that right? Who hasn't? Why not? What recommendations can we give?
- Challenges and questions for the next 10 years
  11. How can we develop or support legal mechanisms that do not require people to constantly read and make decisions?
  12. Could trust models be utilised to help remove individual consent responsibility? For example, trust in AI agents, data cooperatives, and experts.
  13. Would data intermediaries help or exacerbate the situation?
  14. How can we modify the incentive structure to encourage service providers to adopt a different consent model?

4.4.6.2 Requirements for consent/control in different contexts

- Immediate research questions
  1. What do we understand by "context" of privacy decisions? What is all the relevant information to a decision?
  2. How might we gather and organize the consequences of data disclosures that we know? How might we leverage DPIAs to this end? Which existing taxonomy can be adapted for which purpose?
  3. Beyond consequences, how might we incorporate user-centred factors into our understanding of context, such as values, personality, motivation, health status, etc.?
  4. How might we design a *Consent, Control and Communication Manager* 3CM based on these data?
  5. Why did broadly similar concepts such as P3P fail?
  6. What kind of affirmative actions ("unambiguous indication of one's will") might be imagined and implemented in novel modalities such as VR/AR/voice?
- Challenges and questions for the next 3 years
  7. How does the threshold for being informed change depending on context when making privacy decisions? (e.g. do low stakes decisions require less information to be provided?)
  8. How might a privacy decision agent consider context beyond consequences? What kind of resources would be needed to build it and maintain it?
  9. How might we reimagine consent and control for different interaction modalities, such as mixed reality or voice interfaces?
  10. At which moment(s) should consent be asked in different contexts? Should it be disruptive or not?
  11. How should the consequences be communicated, and how can their (long-term and short-term) effects on users' decisions be measured? How might we counteract the potential biases of positive and negative consequences in the database?
  12. How can we responsibly facilitate reflection on privacy decisions?
- Challenges and questions for the next 10 years
  13. How could novel consent models be developed to help positively shape/influence data use and data sharing practices of organisations?
  14. How might we re-write our trained habits that are rooted in the current notice and consent model (e.g., to click accept without thinking)?
  15. How should guidelines and best practices be formulated to meaningfully guide developers/designers to select the best timing, modality, channel etc based on the context?

4.4.6.3 AI-driven (personalized) privacy assistants

- Immediate research questions
  1. What are the knowns and the unknowns of developing such an AI agent?
  2. Which legal standing does a browser extension making cookie decisions on behalf of the user have? Does that apply to other types of technologies?
  3. How do changing ways of interacting with digital technology change the way that people engage with consent and control? (e.g. smartphone, messaging, AI)
  4. What would happen if only data that is needed for the basic functionality of the service (if data minimization was really enforced) was collected, both in terms of benefits and limitations? How would a world without personalization look like?

- Challenges and questions next 3 years:
  5. When/if to ask users if they want to reconsider their decisions, potentially leveraging the teachable moments?
  6. How might an AI agent learn from your decisions as you make them, as well as considering the wider context (for example, new laws or regulations) that might impact sharing decisions?
  7. To what extent can an AI agent make decisions on your behalf? How does it decide whether it can or where it cannot?
  8. What would be different from a human assistant? Should it learn through a process of personalization, and if so, what kind of safeguards should be in place to protect user privacy? What kind of errors would we be willing to tolerate?
  9. Should it learn based on other people's preferences or decisions? If so, what should it learn from others' preferences?
  10. What do we mean by "privacy-friendly' agent? How would we implement that?
  11. Would privacy-enhancing technologies make it entirely unnecessary for us to manage our data and our privacy?
- Challenges and questions for the next 10 years
  12. Does the AI agent always need to make inferences and communicate them to the user? How should a system decide on whether those decisions should be taken by humans or by the agent?
  13. How does consent and control look in imagined futures? For example, if we are not using web portals anymore, but just using AI agents who use web portals on our behalf.
  14. How would the tradeoffs between robustness, accuracy, fairness and privacy be decided and implemented in such a system [20]?
  15. Who would / should develop such systems?
  16. How will different agents interact?

References

1. Polona Car and Filippo Cassetti. 2025. Regulating dark patterns in the EU: Towards digital fairness. *Prepared by the European Parliamentary Research Service for the European Parliament.*
2. Ewa Luger, Stuart Moran, and Tom Rodden. 2013. Consent for all: revealing the hidden complexity of terms and conditions. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). Association for Computing Machinery, New York, NY, USA, 2687–2696. https://doi.org/10.1145/2470654.2481371*
3. Michael Veale and Frederik Zuiderveen Borgesius. 2021. Adtech and Real-Time Bidding under European Data Protection Law. *In German Law Journal, Available at SSRN: https://ssrn.com/abstract=3896855*
4. Acquisti, A., Adjerid, I., & Brandimarte, L. (2013). Gone in 15 seconds: The limits of privacy transparency and control. IEEE Security & Privacy, 11(4), 72-74.
5. Schaub, F., Balebako, R. & Cranor, L. Designing effective privacy notices and controls. *IEEE Internet Computing*. **21**, 70-77 (2017)
6. Schaub, F., Balebako, R., Durity, A. & Cranor, L. A design space for effective privacy notices. *Eleventh Symposium On Usable Privacy And Security (SOUPS 2015)*. pp. 1-17 (2015)
7. Nouwens, M., Liccardi, I., Veale, M., Karger, D. & Kagal, L. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. *Proceedings Of The 2020 CHI Conference On Human Factors In Computing Systems*. pp. 1-13 (2020)

8 Bauer, J., Bergstrøm, R. & Foss-Madsen, R. Are you sure, you want a cookie?–The effects of choice architecture on users' decisions about sharing private online data. *Computers In Human Behavior*. **120** pp. 106729 (2021)

9 Böhme, R. & Köpsell, S. Trained to accept? A field experiment on consent dialogs. *Proceedings Of The SIGCHI Conference On Human Factors In Computing Systems*. pp. 2403-2406 (2010)

10 Turow, J., Hennessy, M., & Draper, N. (2015). The tradeoff fallacy: How marketers are misrepresenting American consumers and opening them up to exploitation. Available at SSRN 2820060.

11 Kelley, P., Cesca, L., Bresee, J. & Cranor, L. Standardizing privacy notices: an online study of the nutrition label approach. *Proceedings Of The SIGCHI Conference On Human Factors In Computing Systems*. pp. 1573-1582 (2010)

12 Gluck, J., Schaub, F., Friedman, A., Habib, H., Sadeh, N., Cranor, L. & Agarwal, Y. How short is too short? Implications of length and framing on the effectiveness of privacy notices. *Twelfth Symposium On Usable Privacy And Security (SOUPS 2016)*. pp. 321-340 (2016)

13 Harbach, M., Hettig, M., Weber, S. & Smith, M. Using personal examples to improve risk communication for security & privacy decisions. *Proceedings Of The SIGCHI Conference On Human Factors In Computing Systems*. pp. 2647-2656 (2014)

14 Wogalter, M., Conzola, V. & Smith-Jackson, T. Research-based guidelines for warning design and evaluation. *Applied Ergonomics*. **33**, 219-230 (2002)

15 Cranor, L., Guduru, P. & Arjula, M. User interfaces for privacy agents. *ACM Transactions On Computer-Human Interaction (TOCHI)*. **13**, 135-178 (2006)

16 Holtz, L., Zwingelberg, H. & Hansen, M. Privacy policy icons. *Privacy And Identity Management For Life*. pp. 279-285 (2011)

17 Tabassum, M., Alqhatani, A., Aldossari, M. & Richter Lipford, H. Increasing user attention with a comic-based policy. *Proceedings Of The 2018 CHI Conference On Human Factors In Computing Systems*. pp. 1-6 (2018)

18 Karegar, F., Pettersson, J. & Fischer-Hübner, S. The dilemma of user engagement in privacy notices: Effects of interaction modes and habituation on user attention. *ACM Transactions On Privacy And Security (TOPS)*. **23**, 1-38 (2020)

19 Zinaida Benenson, Delphine Christin, Alexander De Luca, Simone Fischer-Hübner, Thomas Heimann, Joachim Meyer. Consequence-based Privacy Decisions: a New Way to Better Privacy Management. *In: Alessandro Acquisti, Ioannis Krontiris, Marc Langheinrich, and Martina Angela Sasse. 'My Life, Shared' - Trust and Privacy in the Age of Ubiquitous Experience Sharing (Dagstuhl Seminar 13312). In Dagstuhl Reports, Volume 3, Issue 7, pp. 74-107, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2013) https://doi.org/10.4230/DagRep.3.7.74*

20 Alex Gittens and Bülent Yener and Moti Yung. An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML. *In IEEE Access, Volume 10, 2022. https://doi.org/10.1109/ACCESS.2022.3218715*

4.5 Collective Privacy

Heather Richter Lipford (University of North Carolina at Charlotte, US, heather.lipford@uncc.edu)
Nina Gerber (Technical University of Darmstadt, DE, n.gerber@psychologie.tu-darmstadt.de)
Karola Marky (Ruhr University Bochum, DE, karola.marky@rub.de)
Jessica Vitak (University of Maryland – College Park, US, jvitak@umd.edu)
Camille Cobb (University of Illinois – Urbana-Champaign, US)

4.5.1 Introduction

Privacy has long been conceptualized at the individual level (e.g., [2, 36]). However, with the rise of social and mobile media, as well as the widespread data collection of digital trace data at scale, it is critical to move beyond individual considerations to focus on the privacy needs and risks at the group, organizational, and societal level. For example, in groups or dyads, individuals need to account for and navigate varying privacy preferences and norms; significant research has explored this as it relates to social media (e.g., [15]). Privacy in organizations is likely influenced by the more formalized and hierarchical structures that restrict information flows and create power imbalances between employers and employees. Likewise, technology developers yield significant power in determining the extent to which privacy can–or cannot–be enacted. Finally, privacy evaluations at the societal level run into challenges due to different regulatory landscapes as well as different values and norms.

Beyond these, privacy discussions must now account for the interdependent nature of data creation, use, and ownership. However, accounting for the various entanglements of people and data in a networked world is incredibly challenging and inherently interdisciplinary. Researchers have begun addressing these complexities, with work spanning the computational (e.g., [26]) and social (e.g., [3]) sciences as well as legal scholarship (e.g., [33, 34]).

Researchers have proposed various terms to describe this process, including interdependent privacy [4], networked privacy (e.g., [6]), collective privacy [16], privacy as contextual integrity [24], and comparative privacy [17]; however, these definitions do not fully encompass the complexities involved in a world where both social and technical aspects of data disclosure, collection, and use are constantly evolving and vary based on cultural, community, and contextual factors. Technical solutions alone cannot account for these factors; thus, a more human-centered approach is necessary for imagining and implementing privacy-enhancing solutions for the design of, communication about, and regulation of technologies.

Thus, we propose the following privacy framework to account for these factors:

Socio-collective privacy refers to the idea that privacy is co-constructed and negotiated within social relationships, groups, and society at large. Socio-collective privacy involves the interplay between individual privacy choices and the influence of others – peers, communities, institutions, and socio-cultural norms – in shaping those choices. From this perspective, privacy is not just a personal issue but a shared, social process, embedded in group behavior and societal expectations.

In this working group report, we describe a research roadmap for six aspects of privacy that emerged from the group's conversations at the seminar. Over multiple days, we identified these six clusters by first reviewing and discussing ideas raised during a brainstorming session with all seminar attendees. We discussed intersections and overlaps across these notes, and then expanded into related areas. We decided to focus on three levels beyond the individual

– group, organizational, and societal – and began identifying the areas that had pressing research questions at one or more of these levels. From this discussion, we see this line of research as addressing two overarching goals:

1. understanding how the interplay between the dynamics of and across levels – individual, group, organizational, and societal – impacts key privacy outcomes (e.g., perceptions and mental models; awareness and knowledge; decisions and behaviors); and
2. supporting communities of people to interact and negotiate around privacy.

In the following sections, we describe each of the six research clusters, including their core challenges, a brief summary of the state of the art, and a set of research questions to address that topic at the group, organizational, and/or societal level. We then share a roadmap that outlines research priorities over the short-, medium-, and long-term.

4.5.2 Research Clusters

4.5.2.1 The Role of Collective Actions and Structures

In the discussion, we found that in digitization, the intersection of using digital systems, privacy, and societal values [9] should become a more critical area of study. As digital services increasingly function as societal infrastructure utilized by various kinds of collectives – such as private groups and organizations – we consider it essential to investigate how collective opinions and experiences influence the design of future digital services and their privacy aspects. This includes to what extent the services need to capture and process data, but also the way in which individuals interact with privacy mechanisms (e.g., to manage consent [5]). Based on that, we recognize the role of collectives and their input in shaping digital systems and propose the following overarching research question:

- How can collective opinions/societies shape the design of digital products (inc. privacy aspects)?

Furthermore, communication structures within collectives and societies can serve as vital infrastructures for diffusing essential privacy-related knowledge. This diffusion can occur on multiple levels. For instance, privacy adepts can support their peers in private settings [11]. In doing so, privacy-related knowledge, or at least better privacy settings, are improved in informal contexts. Such support may take the form of "digital housekeeping" [35], which involves individuals helping each other manage their digital privacy practices. Existing research indicates that while these informal support systems are emerging, several barriers exist on the social level [11] and the available technological infrastructure for support is limited [12]. Beyond the private context, it is crucial to examine the privacy practices within organizations. A deeper investigation into these practices can help identify win-win situations that benefit both the organization and its employees, fostering a culture of privacy that enhances trust and collaboration. Here, we propose the following research questions:

- How can privacy-related knowledge be diffused throughout a society/collective?
- How can we create win-win situations to promote privacy (e.g., in organizations)?

Several domains have successfully leveraged collectives and societal structures through existing infrastructures, such as public healthcare initiatives, for instance, aimed at anti-smoking efforts [37] and the automotive ecosystem, which includes manufacturers, repair shops, and official inspection centers. These domains have demonstrated effectiveness in promoting collective well-being and safety in the analog world. Their success prompts an important question: how can we translate these effective strategies into digital systems? We argue that understanding the mechanisms that have facilitated collective action in these

established domains could provide valuable insights for fostering similar initiatives in digital environments. By examining the principles and practices that underpin these successful analog efforts, we can explore ways to adapt and implement them in the context of digital privacy. Based on that, we propose:

- What can we learn from other successful domains (e.g., health) to create a societal infrastructure for privacy and security?

Finally, investigating the research questions detailed above goes together with further research challenges, as it remains unclear how to effectively identify collective measures, establish specific collective goals, and track their progression from a methodological perspective. We argue that addressing these challenges is needed for ultimately developing frameworks that can guide collective action in privacy initiatives as then, we can better understand how to "mobilize" communities and assess the impact of their efforts in promoting privacy in digital environments. Based on that, we propose the following research questions:

- What are collective measures and goals that we can aim for?
- How can we collectively track progress on privacy?

4.5.2.2 Social Norms and Influences

Many theories of privacy (e.g., contextual integrity and boundary regulation (cf. [24, 27]) integrate norm-based conceptualizations of privacy. These existing theories recognize that norms may differ between groups (or organizations, cultures, etc.), and researchers have sought to understand and describe existing norms. Law and policy also regularly reference norms (or "expectations"). Following somewhat directly from these existing ideas, we call out the importance of continuing to pursue a systematic understanding of privacy norms in various groups:

- What are existing privacy norms?
- How do social norms shape/influence privacy behaviors in different contexts?

Understanding the norms of certain groups may be especially relevant to informing conversations about privacy. For example, WEIRD groups have historically had disproportionate influence on the development and design of technologies and their (lack of) privacy affordances. As we return to below, these privacy norms are important to understand in part because of their (likely) influences on norms within other groups.

However, since these societal power structures have also led to knowledge production (i.e., research) disproportionately originating from similar groups, our understanding of their privacy norms already goes beyond our understanding of norms within more niche or marginalized groups. Thus, researchers should also prioritize exploring privacy norms in groups that have been less studied. Additionally, McDonald and Forte [20] argued that taking norm-based perspectives systematically excludes the privacy preferences and needs of individuals who do not fit norms – which is often connected to them being marginalized or vulnerable. Thus, researchers pursuing studies about norms should be sensitive to the impact of vulnerability and consider divergence from the norm, even within studies on non-WEIRD groups. Researchers focused on norms must, therefore, figure out:

- How can we integrate the perspectives of diverse users (e.g. fundamentalists, vulnerable people, children) in research on socio-collective privacy?

Beyond understanding what norms exist, we must also deepen our understanding of how these norms impact individuals' privacy-related behaviors:

- How is individual behavior influenced by groups', organization's and societal norms and behaviors?

- What are the mechanisms in group dynamics that influence individual user behaviors?
- What are the differing effects for different individuals, and from different types of groups?
- What aspects of privacy norms discourage conversations about privacy?
- To what extent is technology design consistent with norms?
- How can collective opinions/societies shape design? Or, how can design reflect collective norms and opinions?

Recognizing that norms can and do shift over time and differ between groups, it is important to understand how privacy-related norms are formed and, in particular, how anti-privacy (or privacy-indifferent) norms come to be. We propose asking:

- What blocks the formation of social privacy norms?
- How do individuals – including those who already exist within normative societal bounds as well as ones at the margins or who have more fundamentalist privacy perspectives – influence groups, organizations, or societies?
- Considering sub-groups whose privacy norms differ from broader societal norms, to what extent are these groups formed around shared beliefs about privacy vs. swayed by some aspect of group dynamics toward a new norm?

Social influences seem to be shifting us toward being less concerned with privacy, so influence can be both a barrier and a potential solution to improving privacy. Perhaps a better understanding of norm formation could empower more marginalized or vulnerable group members to have more meaningful influence on the norms in the groups they belong to:

- What efforts does it take for individuals to have a meaningful influence on groups, organizations, or societies?
- What and how can public influencers (e.g. creators on TikTok or YouTube) have an impact on privacy norms and outcomes?
- How can we design such systems to support meaningful social influences regarding privacy norms?

Finally, we acknowledge that our positionality as privacy researchers comes with an interest in improving privacy and shifting the norm within societies and groups to which we belong to be more privacy-positive. To do this, we must pursue research that informs us about how to achieve these goals. For example:

- How can we create/shape/shift social privacy norms?
- How can we foster societal conversations regarding privacy?
- Can we evaluate the efficacy of focused, short-term privacy campaigns (e.g., against privacy-invasive legislation) compared to sustained, long-term efforts?

4.5.2.3 Privacy Conversations

Social influences such as stories have been found to be an effective trigger for prompting secure and privacy-conscious behavior [29, 28]. In the security context, for example, anti-phishing training that relies on discussions and role-playing have been found to be more effective in increasing self-efficacy and support-seeking than traditional training approaches [7]. Still, privacy especially is a topic that rarely comes up in conversations among non-expert users [8], and is even considered a social taboo by some people [10]. Even experts are hesitant to bring the topic up towards non-experts due to fear of disinterest and negative reactions [11]. Interestingly, in a recent study with a representative U.S. sample most people reported to be interested in having privacy conversations, but felt that other people did not care enough

about the topic, and lacked natural conversation starters [10]. These findings indicate that to fully leverage the potential of social dynamics through privacy conversations, we first have to understand why people decide to (not) engage in conversations about privacy, and, based on the results, identify strategies to facilitate privacy discussions.

Based on these considerations, the following key research questions emerge:

- What are the barriers to privacy conversations?
- How can we trigger privacy conversations?
- How can we foster ongoing privacy conversations?
- What kinds of conversations are beneficial? How are they beneficial?
- What privacy-related topics do people want to talk about?

There has been some initial research on the reasons and effects of conversations mainly focusing on the cybersecurity field. For example, Das et al. [8] conducted an interview study and identified the intention to warn others, share protection strategies, and seek advice were the main motivators of starting security conversations. Their results further imply that people might be hesitant to bring up privacy and security topics since they fear being considered as paranoid, socially inappropriate, or preachy. These findings have later been confirmed for expert users, who also questioned their moral authority to comment upon other people's privacy decisions [11]. Rader et al. [29], replicated by Pfeffer et al. [28], investigated the effect of security stories, finding that stories told in a private context were more likely to drive behavior change, while stories told by security experts were more likely to be retold. In a subsequent analysis, Rader and Wash [30] found that when telling security stories, experts tend to focus on attack mechanisms and prevention measures, while non-experts are more interested in who executed the attack and for which reasons. Further, conversations among developers on platforms such as Stack Overflow have been found to increase knowledge and awareness [14].

Other fields have been studying conversation starters in general, showing that user interfaces can successfully trigger conversations and engagement. For example, public displays have been utilized as conversation triggers, and have been found particularly effective in waiting and non-time-critical situations [1, 23]. Other approaches created interesting nuggets of information from contextual factors, e.g., 'the current temperature is equal to the coldest temperature ever measured in Sao Paulo [21]. Recently, Murtezaj et al. [22] have highlighted the potential of using public security user interfaces for increasing awareness and promoting behavior change.

4.5.2.4 Digital Literacy

One of the goals of privacy interventions is to contribute to the awareness and knowledge of individuals. Digital literacy is also a precursor to making informed privacy decisions and adopting privacy tools. As the above section discussed, one of the ways that users learn about security and privacy is through informal stories shared among friends and family. Organizations and societies also spread knowledge through a variety of formal and informal channels, such as media, communications, influencers, and so on. Thus, understanding and improving digital literacy requires a focus on communities and how knowledge is spread within and through them. Research questions then arise regarding:

- What level of knowledge is sufficient (for individuals or the group)?
- What "digital survival skills" should groups of people have?
- How can that knowledge be diffused throughout a group/community/society?

- To what extent is it important for this knowledge to be spread throughout a community (e.g., all individuals have knowledge) vs. concentrated amongst experts within the group?
- What is the impact of various forms of media on knowledge and awareness?

There has been some research demonstrating what and how users learn about security from others, such as personal contacts [29] and the media [31]. Yet, few have examined how to measure privacy-related knowledge of communities of people, and how that knowledge can impact individual decisions and outcomes. As mentioned previously, a starting point may be to draw upon research strategies or results in other domains, such as public health, where similar issues of community-oriented awareness and knowledge diffusion have been examined. Similarly, few privacy researchers have investigated interventions to promote knowledge sharing and diffusion within groups as a step towards improving privacy management.

4.5.2.5 Cultural Differences

Privacy values – and consequently privacy regulations – vary across the world. This can perhaps most clearly be seen in the different regulatory approaches taken by the European Union (EU) and the United States. The EU enacted the General Data Protection Regulation (GDPR) nearly a decade ago, specifying a range of privacy protections for citizens and requirements for companies that collect personal data. The U.S. has yet to pass comprehensive federal privacy reforms, relying instead on sector-specific and state laws. Thus, it is important to consider the role that cultural differences play in the design, implementation, and (non)use of technology. However, as Li [13] notes, few privacy studies account for cultural differences in their design. Therefore, we suggest three research questions to guide future work in this space:

- How are the cultural differences in privacy in the offline world related to cultural differences in the online world?
- How are cultural differences related to all of the other research we have outlined in this report?
- What aspects of an organizational privacy culture are related to desired privacy behaviors?

4.5.2.6 Bystander Privacy

An important consideration in privacy is that there are people impacted beyond an individual user. For example, many smart devices are capable of capturing information about "bystanders" – those near the device but who have no interaction with or control of that device [18]. A classic example are always-on doorbell cameras that capture people walking by on a public street, who likely have no awareness of when and by whom they are being recorded. Other example include individuals sharing information about other people, such as a person posting a photo of others, or visiting a friend who has smart home devices [19]. Even when individuals are carefully considering privacy, they may be unaware of the "collateral damage" that their technology usage or decisions can cause. For example, people who have utilized DNA services may inadvertently impact the privacy of current and future relatives. We note that the word "bystander" does not adequately describe many of these situations, and other terms have also been used, such as incidental user or non-user [18]. But what unites all of these is the lack of agency and control, or even awareness, of the way that information is collected, used, and shared about oneself by others. Key research questions related to this area are:

- When and how are people impacted by decisions of others and hence lack agency and/or control?

- How can we empower people in multi-user settings and facilitate privacy negotiations?
- How can we leverage protection from collateral damage caused by other individuals' or groups' privacy decisions?

The problem of bystanders has been widely researched and discussed in relation to particular technologies, such as in the examples mentioned above of photo sharing, smart homes [18, 19], and DNA data sharing [32, 25]. General guidelines have been presented in such research, such as to minimize data collection or increase the transparency of data collection. There have also been a range of specific design interventions proposed, yet many of these are specific to a particular device and use case. For example, there has been extensive research on ways to automatically detect and protect the privacy of bystanders in photos. Yet there has only been limited industry uptake of a few of these, such as visual indicators of recording on smart devices as individuals might not buy too obvious devices [19]. Outside of these examples, exploration and evaluation of solutions is still quite limited, and does not cover the range of privacy harms that can come to bystanders in a variety of contexts. Finally, technical solutions may not always be feasible and instead groups may need to rely on social conventions and conversations to negotiate privacy protections – other research areas discussed in this report.

4.5.3 Research Roadmap

We already have many of the methods and tools to approach the questions discussed above, and we propose that we should start pursuing those right away. These methods, which we already use, largely came out of collaborations with psychologists, who tend to take an individual-centric approach to questions and methods. Given that the questions we map here are societal/group, we recognize that our existing interdisciplinary connections will need to expand. So, to deepen and expand the questions in this document and envision next steps, we will need to strengthen collaboration with scholars from other fields such as sociology and anthropology, which will require new types of incentives and funding to support that. These collaborations will be slow to come to fruition, and thus should also be started immediately.

While researchers may already be using relevant methods, there is a gap when it comes to measures. There are fewer existing measures of group-level privacy used in our research community. For example, an open research question is how to measure privacy outcomes at a societal level. Foundations for such measures may again be inspired from other fields that research organizational or societal phenomena. Another challenge is that creating interventions suffers from many of the same challenges as any multi-user system, with more complex design requirements, implementations, and evaluation protocols.

Funding for this research needs to support overcoming these challenges, and incentivize researchers towards our overarching goals. Funding agencies can encourage interdisciplinary work through specialized programs, post-doctoral opportunities, interdisciplinary symposia, summer schools, and other novel structures to encourage researchers from different communities to share ideas and collaborate on projects.

A goal of all of the above research is to provide practical and actionable guidance for socio-technical designs that empower people to achieve their own and their community's privacy needs. Thus, we also discussed the following two questions for the research community:

- How do we help researchers to influence technology design and policy making? In other words, how can we help researchers to communicate and frame results for interested communities, such as regulators, legislators, and technology creators?
- How can we help people choose research questions and study designs that speak to those different audiences?

Finally, we note that our research can also help communities of users contribute to technology and societal outcomes – empowering people to influence designers and regulators towards meetings privacy needs. Thus, we end with two final research questions:

- What privacy advocacy has an impact on regulations?
- How can we support and promote privacy advocacy?

References

1 Alt, F., Kubitza, T., Bial, D., Zaidan, F., Ortel, M., Zurmaar, B., Lewen, T., Sahami Shirazi, A., & Schmidt, A. (2011). Digifieds: insights into deploying digital public notice areas in the wild. In Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (MUM '11). Association for Computing Machinery, New York, NY, USA, 165–174. https://doi.org/10.1145/2107596.2107618

2 Altman, I. (1975). The environment and social behavior: privacy, personal space, territory, and crowding.

3 Anthony, D., Campos-Castillo, C., & Horne, C. (2017). Toward a sociology of privacy. Annual review of sociology, 43(1), 249-269. https://doi.org/10.1146/annurev-soc-060116-053643

4 Biczók, G., & Chia, P. H. (2013). Interdependent privacy: Let me share your data. In Financial Cryptography and Data Security: 17th International Conference, FC 2013, Okinawa, Japan, April 1-5, 2013, Revised Selected Papers 17 (pp. 338-353). Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-39884-1_29

5 Böhme, R., & Köpsell, S. (2010). Trained to accept? A field experiment on consent dialogs. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 2403-2406).

6 Boyd, D. (2012). Networked privacy. Surveillance & society, 10(3/4), p.348.

7 Chen, X., Sacré, M., Lenzini, G., Greiff, S., Distler, V. & Sergeeva, A. (2024). The Effects of Group Discussion and Role-playing Training on Self-efficacy, Support-seeking, and Reporting Phishing Emails: Evidence from a Mixed-design Experiment. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 829, 1–21. https://doi.org/10.1145/3613904.3641943

8 Das, S., Hyun-Jin Kim, T., Dabbish, L.A., and Hong, J.I. (2014). The effect of social influence on security sensitivity. In 10th Symposium On Usable Privacy and Security (SOUPS 2014), pages 143–157, Menlo Park, CA, July 2014. USENIX Association.

9 Friedman, B. (1996). Value-sensitive design. interactions, 3(6), 16-23.

10 Gerber, N., Zimmermann, V., von Preuschen, A., & Renaud, K. (2025). Unpacking the Social and Emotional Dimensions of Security and Privacy User Engagement. In 21st Symposium on Usable Privacy and Security (SOUPS 2025).

11 Nina Gerber and Karola Marky. The nerd factor: The potential of S&P adepts to serve as a social resource in the user's quest for more secure and Privacy-Preserving behavior. In Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022), pages 57–76, Boston, MA, August 2022. USENIX Association. https://www.usenix.org/conference/soups2022/presentation/gerber

12 Horst, H., & Sinanan, J. (2021). Digital housekeeping: Living with data. New Media & Society, 23(4), 834-852.

13 Li, Y. (2022). Cross-Cultural Privacy Differences. In: Knijnenburg, B.P., Page, X., Wisniewski, P., Lipford, H.R., Proferes, N., Romano, J. (eds) Modern Socio-Technical Perspectives on Privacy. Springer, Cham. https://doi.org/10.1007/978-3-030-82786-1_12

14 Tamara Lopez, Thein Tun, Arosha Bandara, Levine Mark, Bashar Nuseibeh, and Helen Sharp. An anatomy of security conversations in stack overflow. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS), pages 31–40. IEEE, 2019. https://doi.org/10.1109/ICSE-SEIS.2019.00012

15 Mansour, A., & Francke, H. (2021). Collective privacy management practices: A study of privacy strategies and risks in a private Facebook group. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-27. https://doi.org/10.1145/3479504

16 Mantelero, A. (2017). From Group Privacy to Collective Privacy: Towards a New Dimension of Privacy and Data Protection in the Big Data Era. In: Taylor, L., Floridi, L., van der Sloot, B. (eds) Group Privacy. Philosophical Studies Series, vol 126. Springer, Cham. https://doi.org/10.1007/978-3-319-46608-8_8

17 Masur, P. K., Epstein, D., Quinn, K., Wilhelm, C., Baruh, L., & Lutz, C. (2025). Comparative privacy research: Literature review, framework, and research agenda. The Information Society, 1-22. https://doi.org/10.1080/01972243.2025.2451863

18 Marky, K., Voit, A., Stöver, A., Kunze, K., Schröder, S., & Mühlhäuser, M. (2020, October). "I don't know how to protect myself": Understanding Privacy Perceptions Resulting from the Presence of Bystanders in Smart Environments. In Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (pp. 1-11).

19 Marky, K., Gerber, N., Pelzer, M. G., Khamis, M., & Mühlhäuser, M. (2022). "You offer privacy like you offer tea": Investigating mechanisms for improving guest privacy in IoT-equipped households. Proceedings on Privacy Enhancing Technologies.

20 McDonald, N., & Forte, A. (2020, April). The politics of privacy theories: Moving from norms to vulnerabilities. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

21 Memarovic, N., Elhart, I., & Langheinrich, M. (2011, December). FunSquare: First experiences with autopoiesic content. In Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (pp. 175-184).https://doi.org/10.1145/2107596.2107619

22 Murtezaj, D., Paneva, V., Distler, V., & Alt, F. (2024, September). Public Security User Interfaces: Supporting Spontaneous Engagement with IT Security. In Proceedings of the New Security Paradigms Workshop (pp. 56-70).

23 Müller, J., Walter, R., Bailly, G., Nischt, M., & Alt, F. (2012, May). Looking glass: a field study on noticing interactivity of a shop window. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 297-306).

24 Nissenbaum, H. (2009). Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford Law Books.

25 Niu, Y., Meng-Schneider, N., Qiu, W., & Kokciyan, N. (2025, April). "I am not the primary focus"-Understanding the Perspectives of Bystanders in Photos Shared Online. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-23).

26 Palen, L., & Dourish, P. (2003, April). Unpacking "privacy" for a networked world. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 129-136).

27 Petronio, S. (2002). Boundaries of privacy: Dialectics of disclosure. SUNY Press.

28 Pfeffer, K., Mai, A., Weippl, E., Rader, E., & Krombholz, K. (2022). Replication: Stories as informal lessons about security. In Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022) (pp. 1-18).

29 Rader, E., Wash, R., & Brooks, B. (2012, July). Stories as informal lessons about security. In Proceedings of the Eighth Symposium on Usable Privacy and Security (pp. 1-17).

30 Rader, E., & Wash, R. (2015). Identifying patterns in informal sources of security information. Journal of Cybersecurity, 1(1), 121-144.

31 Raphael, M. M., Kanta, A., Seebonn, R., Dürmuth, M., & Cobb, C. (2024). Batman Hacked My Password: A Subtitle-Based Analysis of Password Depiction in Movies. In Twentieth Symposium on Usable Privacy and Security (SOUPS 2024) (pp. 199-218).

32 Saqib, E., He, S., Choy, J., Abu-Salma, R., Such, J., Bernd, J., & Javed, M. (2025). Bystander Privacy in Smart Homes: A Systematic Review of Concerns and Solutions. ACM Transactions on Computer-Human Interaction.

33 Solove, D. J. (2004). The digital person: Technology and privacy in the information age (Vol. 1). NYU Press.

34 Solove, D. J. (2025). Artificial intelligence and privacy. Fla. L. Rev., 77, 1.

35 Tolmie, P., Crabtree, A., Rodden, T., Greenhalgh, C., & Benford, S. (2007, September). Making the home network at home: Digital housekeeping. In ECSCW 2007: Proceedings of the 10th European Conference on Computer-Supported Cooperative Work, Limerick, Ireland, 24-28 September 2007 (pp. 331-350). London: Springer London.

36 Westin, A. F. (1968). Privacy and freedom. Washington and Lee Law Review, 25(1), 166.

37 Wood, W., & Neal, D. T. (2016). Healthy through habit: Interventions for initiating & maintaining health behavior change. Behavioral Science & Policy, 2(1), 71-83.

Participants

- Florian Alt
LMU München, DE
- Zinaida Benenson
Universität Erlangen-Nürnberg, DE
- Bettina Berendt
TU Berlin, DE
- Benjamin Berens
KIT – Karlsruher Institut für Technologie, DE
- Nataliia Bielova
INRIA – Sophia Antipolis, FR
- Camille Cobb
University of Illinois – Urbana-Champaign, US
- Ha Dao
MPI für Informatik – Saarbrücken, DE
- Cori Faklaris
University of North Carolina – Charlotte, US
- Simone Fischer-Hübner
Karlstad University, SE
- Nina Gerber
TU Darmstadt, DE
- Sophie Grimme
OFFIS – Oldenburg, DE
- Andreas Gutmann
Ofcom – London, GB
- Dominik Herrmann
Universität Bamberg, DE
- Adam Jenkins
King's College London, GB
- Bailey Kacsmar
University of Alberta – Edmonton, CA
- Apu Kapadia
Indiana University – Bloomington, US
- Farzaneh Karegar
Karlstad University, SE
- Agnieszka Kitkowska
Jönköping University, SE
- Marc Langheinrich
USI – Lugano, CH
- Karola Marky
Ruhr-Universität Bochum, DE
- Mainack Mondal
Indian Institute of Technology – Kharagpur, IN
- Simran Munot
MPI für Informatik – Saarbrücken, DE
- Alena Naiakshina
Universität Köln, DE
- Sameer Patil
University of Utah – Salt Lake City, US
- Maija Poikela
Charité – Berlin, DE
- Sören Preibusch
BfR – Berlin, DE
- Elissa Redmiles
Georgetown University – Washington, DC, US
- Heather Richter Lipford
University of North Carolina – Charlotte, US
- Arianna Rossi
Sant'Anna School of Advanced Studies – Pisa, IT
- Cristiana Santos
Utrecht University, NL
- Anastasia Sergeeva
University of Luxembourg, LU
- William Seymour
King's College London, GB
- Jose Such
Technical University of Valencia, ES & King's College London, GB
- Jessica Vitak
University of Maryland – College Park, US
- Mark Warner
University College London, GB
- Daricia Wilkinson
Arizona State University – Mesa, US
- Yixin Zou
MPI-SP – Bochum, DE

Report from Dagstuhl Seminar 25271

Policy Modeling and Reasoning in Sociotechnical Systems

Marina De Vos[*1], Nicoletta Fornara[*2], Munindar P. Singh[*3], Leon van der Torre[*4], and Jessica Woodgate[†5]

1 **University of Bath, GB.** cssmdv@bath.ac.uk
2 **USI – Lugano, CH.** nicoletta.fornara@usi.ch
3 **North Carolina State University – Raleigh, US.** mpsingh@ncsu.edu
4 **University of Luxembourg, LU.** leon.vandertorre@uni.lu
5 **University of Bristol, GB.** jessica.woodgate@bristol.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25271 "Policy Modeling and Reasoning in Sociotechnical Systems". This seminar brought together researchers from academia and industry who are interested in studying the intersection between computer science, philosophy, logic, ethics, and law to discuss policy modelling and reasoning in a world where computers and humans need to work together. After lightning talks, two invited talks, and an open space topic gathering activity, we settled on four topics for deeper discussion in working groups, interspersed by primer talks from the various communities. The four topics were: 1) Concepts: What are the underlying aspects of this interdisciplinary field, and can they be defined consistently? 2) Agentic AI: How can we enable agents to interact and reason with human users through large language models? 3) Standardisation: How can we facilitate data sharing and compliance in international work with competing business interests? 4) Coevolution: How can we make sure that sociotechnical systems evolve with the societies they operate in? This report provides the abstracts of the talks, including participants' lightning talks, the two invited talks, and four primers, along with short reports from each working group detailing their discussions, including challenges and future opportunities.

Seminar June 29 – July 4, 2025 – https://www.dagstuhl.de/25271

2012 ACM Subject Classification Computing methodologies → Multi-agent systems; Computing methodologies → Philosophical/theoretical foundations of artificial intelligence; Information systems → World Wide Web

Keywords and phrases Multi-agent Systems, Norms and Values, Policy Modelling, Standardisation

Digital Object Identifier 10.4230/DagRep.15.6.132

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

Policy Modeling and Reasoning in Sociotechnical Systems, *Dagstuhl Reports*, Vol. 15, Issue 6, pp. 132–188
Editors: Marina De Vos, Nicoletta Fornara, Munindar P. Singh, Leon van der Torre, and Jessica Woodgate

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Marina De Vos (University of Bath, GB)
Nicoletta Fornara (USI – Lugano, CH)
Munindar P. Singh (North Carolina State University – Raleigh, US)
Leon van der Torre (University of Luxembourg, LU)

Introduction

A policy is a declarative basis for a decision by an individual or organisation. In computing, policies occupy the fruitful space between (social or legal) norms and (architectural or algorithmic) mechanisms. Specifically, a policy is an explicit knowledge-based machine-processable representation that guides the decision-making of an autonomous party. Previously, policies have been studied in computing from a primarily technical standpoint, e.g., in characterising access control in a database management system.

This Dagstuhl Seminar takes a fresh and comprehensive perspective on policies by viewing them as part of a sociotechnical system (STS) comprising intelligent agents and people. What makes policies particularly attractive in the modern milieu, with the rise of Artificial Intelligence (AI), is that they promise transparency, interpretability, and support for effective explanations as well as opportunities for making improvements to individual agents or the system. Yet major challenges remain in understanding how to (1) model policies (so they express stakeholder needs in emerging problem domains precisely, clearly, and succinctly), (2) provide a formal semantics for them that respects the autonomy of and interactions between individual agents, and (3) reason about them efficiently. Among these challenges are handling conflicts, allowing agents to deviate from policies, and understanding the trade-offs between system architectures that accommodate different levels of autonomy and efficiency.

Organisation of the Seminar

The seminar was held over the week of June 29 to July 04, 2025 (Monday to Friday with arrival on Sunday). We had 36 on-site participants. We started the first day with a two-minute lightning talk of each of the participants. Each participant was asked to answer the following questions: "what I'm doing, what I want to discuss here, and what input I would like". Before the Monday lunch, we had our two invited vision talks:

Matthew Arrott presented his vision from an industrial perspective in "Adoption of Interaction Protocols in Multi-Party Financial Transaction Platforms".

Pınar Yolum gave her view from an academic perspective in "Sociotechnical Thinking of Privacy".

After lunch, we brainstormed in random groups on topics for discussion for the week, followed by an open space session to determine the working groups for the week. We settled on

- Concepts: What are the underlying aspects of this interdisciplinary field and can be defined consistently?
- Agentic AI: How can we make agents interact and reason with human users through large language models (LLMs)?

- Standardisation: How can we facilitate global data sharing and compliance in a world with competing business interests?
- Coevolution: How can we make sure that STS evolve with the societies they operate in?

For the remainder of the seminar, we focused on discussing these topics, with the intention of coming up with plans for research publications in each of them. This was done through breakout groups and plenary discussions for cross-fertilisation between groups. Participants were encouraged to move between groups. At the end of each day, we had one or two primers of one of the participating disciplines. These primers addressed

- Judith Simon, "Ethics".
- Rigo Wenning, "Standardisation".
- Amit K. Chopra, "Programming with Norms".
- Harko Verhagen and Julian Padget, "The Value of Values".
- Beishui Liao, "Logic for New Generation AI: An Argumentation Based Methodology".
- Joris Hulstijn, "Regulatory Supervision".

In keeping with Dagstuhl tradition, we organised a social event. While normally this would be an excursion or a walk, we decided, due to the excessive heat that week, to stay at the Castle and have a local wine tasting. Julian Padget and Harko Verhagen kindly hosted the event. The final session of the seminar was dedicated to looking ahead and discussing plans for future collaboration and publications as a result of working group discussions. We also discussed the possibility of a follow-up Dagstuhl Seminar. Overall, the seminar was highly engaging, intellectually stimulating, and a great success.

Outcome of the Seminar

Scientific content

Open problems: Each of the four working groups selected by the participants focused on advancing understanding of open problems in the field. The working group reports at the end of this document detail the context of each working group, the progress they made during the seminar, which further challenges were identified, and how the group will take the topic forward.

New connections: The seminar brought together researchers who are working in the broad area of policy modelling but who approach that subject from different angles. Through primers, discussion in the streams, and in social time, we saw interesting exchanges of ideas among the subcommunities. We hope that these interactions will be fostered into new collaborations.

Extensive collaborations: Participants extensively discussed problems and challenges with colleagues, establishing new avenues for collaboration or reinforcing existing ones.

2 Table of Contents

3 Research area

Marina De Vos (University of Bath, GB)
Nicoletta Fornara (USI – Lugano, CH)
Munindar P. Singh (North Carolina State University – Raleigh, US)
Leon van der Torre (University of Luxembourg, LU)

Research Challenges

This seminar synthesised research perspectives from computing with insights from the law, public administration, and the social sciences. In particular, the relevant communities in computing include Semantic Web, Knowledge Representation and Reasoning, Logic Programming, Multi-agent Systems, Privacy and Security, and Legal Informatics, which we introduce below.

This seminar investigated the entire lifecycle of problems in policy dealing with STS. The design of a policy in an STS faces a fundamental trade-off between the *autonomy* accorded to member agents and the *control* exercised over those agents to guide them toward stakeholder objectives.

- Architecture. Architecture here encompasses the social and technical tiers of an STS. It captures what assumptions member agents can make about each other and what guarantees they can expect from the social and technical tiers. These guarantees can be expressed as policies and motivate an interest in an expanded view of policies. These guarantees may include organisational controls, such as sanctions applied for deviation from a policy.
- Models. Models concern how policies are conceived, including the languages in which they are expressed and how they relate to other parts of the relevant information systems. In a formal sense, the models reflect the architecture in an information model along with the needs of the domain. Models include considerations of the formal semantics, e.g., in terms of the computations that can be realised from a system and a determination of which computations are compatible with a given set of policies.
- Reasoning. Reasoning concerns how decisions can be derived from policies, given the facts and reasoning about policies, such as whether they conflict or one subsumes another. It incorporates monitoring (for ease of exposition) to enable reasoning on specific instances as well as determining if a particular deviation was legitimate.
- Methodology. Methodology concerns ways in which policies may be specified for an STS, given stakeholder requirements. It incorporates making changes in light of observed decisions, whether deviations took place, and whether the outcomes and the deviations (if any) were deemed legitimate.

The objective of this seminar was to provide a platform for researchers from different fields to form a new community. Specifically, we sought to motivate participants to define new research problems along with promising ways of tackling them.

Contributing Research Fields

Here is an overview of the various fields and communities that study policies in various forms.

Semantic Web. addresses languages and methods for encoding information formally so that intelligent agents can use that information without the risk of ambiguity that attaches to informal notations such as natural language. The Semantic Web is focused on ontologies – a form of expressive metadata – along with algorithms for formal reasoning on the ontologies. A famous realisation of the Semantic Web is Linked Data, wherein data are mutually linked to enhance their meaning and usefulness.

Knowledge representation and reasoning. (KRR) addresses ways to represent information so it can be used by an intelligent agent in its reasoning and planning. KRR incorporates findings from folk psychology to design formalisms that facilitate solving complex tasks. KRR traditionally emphasises computational logic to automate reasoning and includes studies of rules. Here, logic programming concerns models that automatically generate solutions to formally represented problems, thus obviating the need for procedural algorithms. KRR also includes deontic logic, which focuses on reasoning about obligations, permissions, and rights and thus relates closely to policies.

Multi-agent systems. (MAS) addresses developing systems of autonomous agents that are logically decentralised. A MAS is characterised by how its member agents interact, which leads to research into formal communication languages (described by the information they convey and the social relationships they affect) as well as intelligent decision-making about whether and when to perform a communicative act and whether and how to respond to a communicative act by another agent. These topics are thus well-aligned with policies. The connection is stronger in the Normative MAS (NorMAS) subfield, which focuses on social and legal norms and on organisational architectures.

Privacy, security and policies. for their proper handling of specific data are crucial for many areas of research. Namely, the Semantic Web, because it is focused on enabling data-sharing, for the legal domain, because privacy is regulated by data protection regulations, like GDPR, and the database community for its access control studies. Various privacy issues may arise at different stages of information management, from its collection to its processing and dissemination.

Legal informatics. addresses the formal modelling of laws concerning the usage of AI (with respect to AI provider or platform, e.g., with respect to privacy) and any domain where AI is used (e.g., with respect to the liability of a robot or a robot operator). This field relates well to the above areas of computing; it presents them with challenging problems and benefits from their solutions.

Seminar Topics

Legal knowledge representation and reasoning. Laws can be understood as high-level norms on behaviour and interaction in a society. Policies can be understood as operationalisations of laws. Policies in the legal sense are still high-level in that they may not be readily computed with, partly because they are expressed in natural language and partly because they reference information that may not be readily computationally characterised.

This seminar will study formal policy models and concomitant methodologies through which legal nuance can be reliably represented and reasoned about. Important concerns

include (1) conflicts between policies (e.g., due to jurisdiction or other attributes), (2) the (constrained) freedom of an agent to violate a policy, and (3) revisions.

Reasoning about correctness. Policies occupy the space between (legal and social) norms and agent behaviour. In our STS framework, this exposes important challenges concerning (1) validation: whether a policy represents stakeholder needs as evidenced in the applicable norms, (2) verification: whether agent interactions as designed respect the applicable policies, and (3) compliance: whether agent interactions as realised deviate from the applicable policies. Responses to these challenges determine how an STS and its member agents can be improved through continual revision – e.g., deviations may be justified by an "upstream" argument that the computational policy omitted a possibility allowed by the underlying informal policy.

This seminar explored not only policies as artefacts (and how to represent and reason about them computationally) but also the human-driven processes through which they are developed and revised.

Sociotechnical architecture. The above vision calls for new thinking about the architecture of STS. That is, we need to capture what an agent can expect from other agents and from the STS. Specifically, these expectations concern how information and control are distributed: are some policies enforced through technical artefacts (and difficult to violate without circumventing those artefacts)? Are there compliance checks when onboarding new agents into the system? Are interactions monitored? Are there social controls in place, e.g., reputation or eviction? Can sanctions arising from deviations be negotiated? This seminar studied alternative architectures as devised in informal real practice (such as the law and organisations), semiformal practice (such as access control and break-the-glass scenarios in healthcare, and more formal models (such as in organisational models in MAS).

Applications. Policy technologies are a case where engineering has gotten ahead of science. For example, Open Digital Rights Language (ODRL) is a W3C Recommendation, it is a policy expression language that provides an information model and a vocabulary for policies about the usage of digital assets and services. Even though ODRL is gaining traction (it's now in version 2.2), it lacks a formal model and semantics. We anticipate that use cases from ODRL, albeit limited to data policies, could be interesting real-life challenges for our discussions of the policy lifecycle, and especially on the formal models. This seminar will study practical use cases of policies in practice and identify research to give practice a robust foundation so that it can proceed with greater rigor and generality.

4 Overview of Talks

4.1 Prosociality and Ethics in Sociotechnical Systems

Nirav Ajmeri (University of Bristol, GB)

Prosociality refers to voluntary actions or behaviours intended to benefit an individual or society at large. Normative ethical principles are philosophical guidelines that define acceptable and unacceptable behaviours, offering a foundation for evaluating actions based on their ethical implications. As AI systems increasingly influence decisions with societal impact, their ability to act in prosocial and ethically aligned ways becomes critical. AI

Agents designed today often prioritise the goals and preferences of their primary users – risking outcomes that reinforce existing privileges and disadvantage vulnerable individuals or marginalised communities. Even agents designed for multi-stakeholder contexts may inadvertently overlook broader societal implications. At this Dagstuhl Seminar, I explore how normative ethics can inform the design of AI systems that account for the well-being of others, and discuss recent methods for embedding ethical norms and prosocial behaviours in agents through interaction and social learning. These methods enable more equitable, fair, and responsible STS – better aligned with the values of all stakeholders.

4.2 Introducing exceptions, accountability, commitments, information protocols into JaCaMo and Jason

Matteo Baldoni (University of Turin, IT) and Cristina Baroglio (University of Turin, IT)

JaCaMo and JADE + 2COMM: we show the benefits of explicitly representing social relationships between agents. This approach improves code modularity and interaction flexibility. Additionally, treating commitments as manipulable resources allows agents to reason about their interactions and strategically decide how and when to engage with others to pursue their own goals, enhancing the overall system's effectiveness.

JaCaMo extended with exceptions and accountability: we aimed to enhance the robustness of MAS. The first extension to JaCaMo introduces an exception handling mechanism tailored for MAS, while the second uses accountability to establish feedback chains among agents. Both extensions offer high-level abstractions and follow a unified approach to support the design of robust MAS capable of functioning correctly despite disruptions.

Jason and his friends Orpheus and Azorus: Orpheus and Azorus offers a programming model designed to enhance commitment-based reasoning in decentralised MAS. It uses declarative specifications centred on commitments and integrates them with information protocols. It supports reasoning about both goals and commitments and unifies three key technologies: Jason (a BDI-based agent programming model), Cupid (a formal language for commitments), and BSPL (a protocol language for information exchange). The model is implemented and shown to effectively represent complex business logic patterns.

4.3 All Intelligent Agents Should Speak a Formal Language

Victor Charpenay (Mines Saint-Étienne, FR)

Modern STS usually involve many more machines than humans. Among each other, machines always speak a formal language, which can be as simple as JSON (or better, JSON-LD) or as elaborate as FIPA-ACL. Because humans are outnumbered in STS, the best way for them to interact with machines is to speak the language of machines, via a dedicated graphical user interface for example. The ability of Transformers to model natural language should not encourage engineers to make machines speak natural language but rather to develop new forms of user interface to enhance the fluency of humans in formal languages.

4.4 Meaning-Based Abstractions for Agentic AI

Amit K. Chopra (Lancaster University, GB)

The Agentic AI paradigm is concerned with creating LLM-powered agents that take actions in the real world on behalf of their users. The promise of LLMs lies in their potential to reduce the knowledge engineering effort needed to build agents. Instead of being explicitly programmed, the agents would exploit LLMs to engage in a natural language dialog with their users, figure out the relevant constraints, and act accordingly. Several software frameworks (including protocols) lay claim to realizing Agentic AI. However, these frameworks miss crucial features about the context of real-world actions and their meanings.

Via the notion of norms, meaning is what much research in MAS has been concerned with. The idea is that communications between agents change the normative state of a system and this state is what matters to agents (and the principals they represent) in their reasoning. Recent work has shown how agents can engage flexibly on the basis of norms. A great direction for research is the synthesis of this body of work with LLM-based reasoning to realise more flexible, practical, and reliable Agentic AI.

4.5 Normative Multiagent Systems

Mehdi Dastani (Utrecht University, NL)

Normative systems are widely recognised as an effective means of regulating agent behaviour in MAS. Since the introduction of new norms alters system behaviour, there is a need for formal methodologies to model such dynamics. One line of research in our research group is to addresses this by treating the addition of norms as system updates and introducing formal update semantics to capture their impact.

Another contribution examines norm revision as a mechanism for improving system performance and ensuring the fulfilment of desirable properties. By analysing revisions such as relaxation and strengthening, and illustrating their effects through practical scenarios, our research explores how adaptive adjustment of norms can align MAS behaviour with system-level objectives.

A complementary approach investigates the challenges of maintaining effective norm enforcement in dynamic environments, where objectives evolve and previously defined norms may lose their effectiveness. To address this, we have introduced the data-driven norm revision framework. This framework automatically synthesises and revises conditional prohibitions with deadlines using system execution data. By analysing behavioural traces, the framework generates revised norms that more accurately distinguish between acceptable and unacceptable behaviours. Empirical evaluation using an advanced urban traffic simulator demonstrates that our approach significantly outperforms original norms in supporting the achievement of system objectives.

Collectively, these research directions advance the theory and practice of dynamic norm management in MAS by providing formal models for norm updates, conceptual tools for norm revision, and data-driven methods for adaptive norm synthesis.

4.6 Engineering Human-AI Teams: Norms, Values, and Responsible Collaboration

Davide Dell'Anna (Utrecht University, NL)

Recent advances in AI have made it necessary or desired for humans to get involved in interactions with AI systems on a daily basis. A key factor for the acceptance and responsible use of AI systems in STS is their ability to understand and adapt to personal, social, and legal norms. My research focuses on methodologies and mechanisms for designing AI systems that collaborate with humans synergistically and proactively as Human-AI teams where members amplify each other's intelligence by combining their complementary strengths [3]. I study how to represent, computationally, human social constructs such as norms, values, and team properties, and how to develop automated adaptive and data-driven mechanisms that ensure that AI behaviour is responsible, trustworthy, and justifiable [2, 1].

References

1 Davide Dell'Anna, Natasha Alechina, Fabiano Dalpiaz, Mehdi Dastani, and Brian Logan. Data-driven revision of conditional norms in multi-agent systems. *J. Artif. Intell. Res.*, 75:1549–1593, 2022. URL: https://doi.org/10.1613/jair.1.13683, doi:10.1613/JAIR.1.13683.

2 Davide Dell'Anna and Anahita Jamshidnejad. SONAR: an adaptive control architecture for social norm aware robots. *Int. J. Soc. Robotics*, 16(9):1969–2000, 2024. URL: https://doi.org/10.1007/s12369-024-01172-8, doi:10.1007/S12369-024-01172-8.

3 Davide Dell'Anna, Pradeep K. Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M. Jonker, Catharine Oertel, and Pınar Yolum. Toward a quality model for hybrid intelligence teams. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 434–443. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024. URL: https://dl.acm.org/doi/10.5555/3635637.3662893, doi:10.5555/3635637.3662893.

4.7 Sociotechnical Systems: Stay relevant: Norm Change and Value-Alignment

Marina De Vos (University of Bath, GB)

Joint work of Marina De Vos, Andreasa Morris-Martin, Julian Padget, Jack McKinlay, Mattias Brännström, Lili Jiang

When implementing autonomous agents in an STS, it is critical that their actions are in line with expected behaviours and with the social values of the stakeholders of the systems. To achieve this, the system and the agents should be equipped to reason about the norms and values of the system and how these can be affected by their actions. For modelling norms within a system we use the institutional action language InstAL which maps computationally to answer set programming. As the STS evolves over time, we allow participants to request norm changes which, if approved, are implemented in the system by updating the norms through inductive logic programming. Recently, we started exploring incorporating values and value-alignment, either through stand-alone value extraction from policy documents using LLMs or value-based decision-making for agents using argumentation frameworks.

4.8 Trust envelopes: Vehicles of History and Destiny

Beatriz Esteves (Ghent University, BE)

The original vision of the Web was one of decentralisation; however, this ideal is not reflected in the current landscape. While Web-based services and data inherently originate from diverse sources, the exchange of data – particularly personal data – is predominantly governed by a limited number of large BigTech companies. This concentration of control has contributed to growing public distrust in these services.

Our argument is not that data flows are absent, but rather that they are inefficient and misaligned with technological, legal and business principles. On one side, companies, uncertain about user retention, engage in aggressive data collection from the very first interaction. On the other, users – seeking the convenience of online services – often accept privacy policies and terms of service without adequate scrutiny.

We hypothesise that meaningful and trustworthy data exchange, conducted at every point in time where a data point needs to be exchanged with a clearly defined purpose, can foster evolving, trust-based relationships between individuals and organisations. To support this vision, we introduce the concept of trust envelopes. A trust envelope serves as a carrier of both the historical context and the intended future use of a data element. By accompanying data with usage policies, provenance and other contextual information, trust envelopes enable recipients to verify the origin and quality of the data and to use it in accordance with the source entity's preferences. As such, they enable well-intentioned actors to engage in responsible data exchange without facing the disproportionate obstacles currently present in the digital ecosystem, while those with less sincere motives will be unable to exploit the advantages of these evolvable trust relationships.

4.9 Modelling and Reasoning about Policies in Sociotechnical Systems

Nicoletta Fornara (USI – Lugano, CH)

The problem of modelling and reasoning on norms, policies, agreements and licenses is increasingly crucial in many fields of application and research, e.g. in the design and development of STS, in the regulation of autonomous agents on the Web of Things (see the Dagstuhl Seminar 23081 Agents on the Web[1]), for the governance of the use and exchange of personal and business knowledge graphs between parties (see Dagstuhl Seminar 25051 Trust and Accountability in Knowledge Graph-Based AI for Self-determination[2]), for the governance of the exchange of Data Spaces, Personal Data Stores (in the Solid open standard) and for the second used of health data. Automatically reasoning on the semantics of policies is crucial for providing different types of services, for example, what-if analysis, access control, monitoring and sanctioning, and conflict detection. In my research I studied the

[1] Dagstuhl Seminar 23081 Agents on the Web (Feb 19 – Feb 24, 2023) `https://www.dagstuhl.de/23081`
[2] Dagstuhl Seminar 25051 Trust and Accountability in Knowledge Graph-Based AI for Self Determination (Jan 26 – Jan 31, 2025) `https://www.dagstuhl.de/25051`

formalisation of frameworks for modelling and reasoning on policies by using Semantic Web Technologies and rule languages. Together with my colleagues, we proposed a model to represent and reason about obligations, prohibitions and permissions by extending the ODRL policy language [1] and the T-NORM model of norms able to regulate classes of actions whose performance is temporally constrained [2]. Since 2021, I have been co-chair of the W3C ODRL (Open Digital Rights Language) Community Group in which I mainly coordinate the activities of the group that defines the semantics of ODRL 2.2[3] [3]. Important challenges are the completion of the definition of the formal and operational semantics of the ODRL 2.2 language and the proposal of a new version of the model to overcome its current limitations, which must pass through the study of actual use cases, including its use in STS, and the definition of the main requirements that the new model should meet [4].

References

1 Fornara, N., & Colombetti, M. (2019). Using semantic web technologies and production rules for reasoning on obligations, permissions, and prohibitions. Ai Communications, 32(4), 319-334. https://doi.org/10.3233/AIC-190617.

2 Fornara, N., Roshankish, S., & Colombetti, M. (2021, May). A framework for automatic monitoring of norms that regulate time constrained actions. In International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems Vol. 13239 (pp. 9-27). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-16617-4_2.

3 Bonatti, P. A., Fornara, N., & Harth, A. (2025). Towards a Formal Semantics of the Open Digital Rights Language (ODRL 2.2). ESWC 2025 Workshops and Tutorials Joint Proceedings. 1st International Workshop on ODRL and beyond: Practical Applications and challenges for poLicy-base access and usage control (OPAL2025). Portorož, Slovenia 1st June 2025. Vol-3977. https://ceur-ws.org/Vol-3977/OPAL2025-4.pdf.

4 Cimmino, A., & Fornara, N. (2025). Improving ODRL 2.2: current limitations and theoretical solutions. ESWC 2025 Workshops and Tutorials Joint Proceedings. 1st International Workshop on ODRL and beyond: Practical Applications and challenges for poLicy-base access and usage control (OPAL2025). Portorož, Slovenia 1st June 2025. Vol-3977. https://ceur-ws.org/Vol-3977/OPAL2025-6.pdf.

4.10 Regulatory Supervision

Joris Hulstijn (Utrecht University, NL)

This tutorial presents a brief summary of some of the theory in public administration, IT auditing and law, that is relevant to regulatory supervision and compliance, especially where it concerns corporate regulation; not individual citizens. In particular, we discuss the topics of responsive regulation, (enforced) self-regulation, system-based supervision (also known as collaborative compliance) and the interpretation of open norms. These notions are illustrated by an example from the customs domain: "towards data-driven supervision". In a data-driven supervision process, regulators rely on data from documents provided by the company being supervised. In that case, the central question is: how often and when should regulators schedule inspections to verify the data?

[3] ODRL Formal Semantics, Draft Community Group Report https://w3c.github.io/odrl/formal-semantics/

4.11 What Do We Need for Software-based Normative MAS?

Timotheus Kampik (SAP Berlin, DE & Umeå University, SE)

Indeed, for technologies such as business rule engines, enforcing and verifying compliance with normative requirements is a primary use-case class. Accordingly, we have "normative" capabilities in software systems that govern sociotechnical MAS. However, the "intelligent" handling of exceptions to norms, the deliberate violation of norms based on values, and the evolution of norms is still up to humans, incurring substantial social efforts in organisations, and causing conflicts between norms and values. Accordingly, more flexible architectures and practically more expressive abstractions are required for moving the operational workload of norm management and evolution from the human to the software agent level.

4.12 Logic for New Generation AI: An Argumentation Based Methodology

Beishui Liao (Zhejiang University, CN), Réka Markovich (University of Luxembourg, LU), Leon van der Torre (University of Luxembourg, LU)

This talk introduces a methodology to address the challenges of managing inherently conflicting and evolving policies, norms, and values within complex STS. It argues that formal argumentation provides the necessary rigorous, structured foundation for representing these concepts, including violations and causal links, and for deriving defensible conclusions. To tackle these complexities, we propose a comprehensive integrated framework built upon formal argumentation. This framework consists of six core, interconnected components: 1) A unified representation using defeasible rules to formalise norms, policies, and their violation conditions, naturally handling exceptions and priorities; 2) A conflict and violation resolution engine based on argumentation theory to systematically identify and adjudicate conflicts and violations through argument evaluation; 3) A dynamic adaptation mechanism using argumentation revision to evolve the system by adding, modifying, or retracting rules and arguments in response to change; 4) A causal attribution interface combining argumentation with causal inference to link normative states (compliance/violation) to root causes of outcomes; 5) An efficient computation strategy employing locality and modularity for scalable reasoning in large systems; 6) A neuro-symbolic integration pathway leveraging LLMs for tasks like natural language parsing and argument generation, while relying on the argumentation core for rigorous, explainable reasoning and validation.

References

1 Michael Anderson, and Susan Leigh Anderson. *Geneth: a general ethical dilemma analyzer*. Paladyn J. Behav. Robotics, 9(1):337–357, 2018.
2 Edmond Awad, Michael Anderson, Susan Leigh Anderson, and Beishui Liao. *An approach for combining ethical principles with public opinion to guide public policy*. Artif. Intell., 287: 103349, 2020.

3 Pietro Baroni, Guido Boella, Federico Cerutti, Massimiliano Giacomin, Leendert W. N. van der Torre, and Serena Villata. *On the input/output behavior of argumentation frameworks*. Artif. Intell., 217:144–197, 2014.

4 Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. *Automatic evaluation of design alternatives with quantitative argumentation*. Argument Comput., 6(1):24–49, 2015.

5 Pietro Baroni, Massimiliano Giacomin, and Beishui Liao. *Locality and Modularity in Abstract Argumentation*. In Handbook of Formal Argumentation, pp. 937-980, 2018.

6 Chen Chen, Pere Pardo, Leendert van der Torre, and Liuwen Yu. *Weakest link in formal argumentation: Lookahead and principle-based analysis*. In Andreas Herzig, Jieting Luo, and Pere Pardo, editors, Logic and Argumentation – 5th International Conference, CLAR 2023, Hangzhou, China, September 10-12, 2023, Proceedings, volume 14156 of Lecture Notes in Computer Science, pages 61–83.

7 Haixiao Chi and Beishui Liao. *A quantitative argumentation-based automated explainable decision system for fake news detection on social media*. Knowledge-Based Systems, 242:108378, 2022.

8 Phan Minh Dung. *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*. Artificial intelligence, 77(2):321–357, 1995.

9 Xiaotong Fang, Zhaoqun Li, Chen Chen, and Beishui Liao. *Llm-aspic+: A neuro-symbolic framework for defeasible reasoning*. To appear, 2025.

10 Bettina Fazzinga, Sergio Flesca, and Francesco Parisi. *On the complexity of probabilistic abstract argumentation frameworks*. ACM Trans. Comput. Log., 16(3):22:1–22:39, 2015.

11 Anthony Hunter. *Some foundations for probabilistic abstract argumentation*. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, Computational Models of Argument – Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012, volume 245 of Frontiers in Artificial Intelligence and Applications, pages 117–128. IOS Press, 2012.

12 Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, Zhichao Jiang. *Causal Inference*. Engineering, Volume 6, Issue 3, March 2020, Pages 253-263.

13 Beishui Liao. *Toward incremental computation of argumentation semantics: A decomposition-based approach*. Ann. Math. Artif. Intell., 67(3-4):319–358, 2013. 10.1007/S10472-013-9364-8.

14 Beishui Liao. *On interdisciplinary studies of a new generation of artificial intelligence and logic*. Social Sciences in China, 43(3):5–19, 2022.

15 Beishui Liao and Huaxin Huang. *ANGLE: an autonomous, normative and guidable agent with changing knowledge*. Inf. Sci., 180(17):3117–3139, 2010.

16 Beishui Liao, Li Jin, and Robert C Koons. *Dynamics of argumentation systems: A division-based method*. Artificial Intelligence, 175(11):1790–1814, 2011.

17 Beishui Liao, Kang Xu, and Huaxin Huang. *Formulating semantics of probabilistic argumentation by characterizing subgraphs: theory and empirical results*. J. Log. Comput., 28(2):305–335, 2018.

18 Beishui Liao, Leendert van der Torre. *Explanation Semantics for Abstract Argumentation*. COMMA 2020: 271-282.

19 Beishui Liao, Michael Anderson, and Susan Leigh Anderson. *Representation, justification, and explanation in a value-driven agent: an argumentation-based approach*. AI Ethics, 1(1): 5–19, 2021.

20 Beishui Liao, Pere Pardo, Marija Slavkovik, and Leendert van der Torre. *The jiminy advisor: Moral agreements among stakeholders based on norms and argumentation*. Journal of Artificial Intelligence Research, 77:737–792, 2023.

21 Beishui Liao, Leender van der Torre. *Attack-defense semantics of argumentation*. COMMA 2024: 133-144.

22 Liuwen Yu and Davide Liga Réka Markovich. *Addressing the right to explanation and the right to challenge through hybrid-ai: Symbolic constraints over large language models via prompt engineering*. In The 20th International Conference on Artificial Intelligence and Law, 2025.

23 Liuwen Yu, Réka Markovich, and Leendert van der Torre. *Interpretations of support among arguments*. pages 194–203. IOS Press, 2020.

24 Liuwen Yu, Dongheng Chen, Lisha Qiao, Yiqi Shen, and Leendert van der Torre. *A principle-based analysis of abstract agent argumentation semantics*. In Meghyn Bienvenu, Gerhard Lakemeyer, and Esra Erdem, editors, Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021, Online event, November 3-12, 2021, pages 629–640, 2021.

25 Liuwen Yu, Mirko Zichichi, Réka Markovich, and Amro Najjar. *Enhancing trust in trust services: Towards an intelligent human-input-based blockchain oracle (ihibo)*. In 55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022, pages 1–10. ScholarSpace, 2022.

26 Liuwen Yu, Leendert van der Torre, and Réka Markovich. *Thirteen challenges in formal and computational argumentation*. In Gabbay, D., Kern-Isberner, G., Simari, G.R., Thimm, M. (eds.) Handbook of Formal Argumentation, pages 890–976. College Publications, 2024.

4.13 Abstract of Research

Réka Markovich (University of Luxembourg, LU)

I research computational legal theory and study its applications in AI and legal reasoning. My focus areas are legal knowledge representation, norMAS, deontic logic, machine ethics, and explainable AI (XAI). Computational legal theory is about reconstructing fundamental legal concepts and structures in a formal language. One of the topical foci of mine has a special relevance for policy modelling and reasoning in STS: I have been investigating the formal structure of normative positions. The theory of normative positions is based on the theory of W.N. Hohfeld, who differentiated between four types of positions often referred to as a "right" (claim-right, privilege/freedom, power, immunity) and their corresponding "duty" positions (duty, no-claim, liability, disability). The agents in these positions are in normative relations with each other. This differentiation and the characterisation of the positions and the relations are crucial in order to avoid terminological mess in the law and any system aiming at implementing it, but also for understanding further fundamental concepts playing an essential role in STS, such as competence, responsibility, authority, commitment. Hence the concepts and their adequate formalisation contribute to build, as Marek Sergot puts it, the "characteristic of all forms of regulated and organised agent interaction".

4.14 Sociotechnical Systems: From Dialogue to Decisions

Pradeep Murukannaiah (TU Delft, NL)

Engineering STS is the overarching theme of my research. I envision an STS as a system that supports rich interactions among principals (humans or organisation) and computational agents, enabling a variety of individual and societal applications. In an STS, principals are paramount. Principals act autonomously (based on values) and are accountable to each other (as specified by norms). Agents, in contrast, support decision-making by principals. In this seminar, I explore how to connect dialogue among stakeholders to decision-support by agents.

4.15 Computational Machine Ethics

Vivek Nallur (University College Dublin, IE)

Increasingly machines will be called upon to take ethically charged decisions. In these situations, it is important for the machine to behave in an ethically acceptable manner. The ability of an ABM-based simulation to "play out a few steps into the future", allows the agent to make a principled decision. Those decisions can be taken in a manner that respects multiple stakeholder values, i.e., in a pro-social manner. The agent also attempts to anticipate the humans around it, by understanding the cognitive model of the human interacting with it, and the various possible biases that impact human decision-making.

4.16 Normative Regulation of the Industry of the Future

Luis Gustavo Nardin (IMT Mines Saint-Étienne, FR)

Modern industry is compelled to become more flexible, adaptable, resilient, sustainable, and human-centred in order to evolve and remain competitive. We claim that these requirements can be fulfilled by coupling industrial processes modelling with normative aspects to define a set of design principles for governing industrial systems to operate trustworthy and sustainably, and to respond quickly and flexibly to exogenous and endogenous changes. The explicit normative representation and reasoning enable agents to both adapt the execution of industrial processes to unexpected situations and conditions, and to transparently and intelligibly express their decisions to an human operator. We aim to create normative regulation mechanisms, design regulation architectures and implement platforms that enable agents to operate in heterogeneous and dynamic industrial settings and reason about normative aspects to enhance flexibility, resilience, trustworthiness, and sustainability for the Industry of the Future.

4.17 CERTAIN: Towards Traceability and Regulatory Compliance in AI

Sebastian Neumaier (FH – St. Pölten, AT)

As AI systems become increasingly embedded in critical sectors, ensuring regulatory compliance and ethical integrity becomes essential. In my talk, I introduce the CERTAIN project (`https://certain-project.eu/`), which aims to develop a comprehensive framework for the traceability and compliance checking of AI systems within the evolving regulatory landscape of the European Union. Central to this effort is a Semantic MLOps Engine and a RegOps Engine:

- The semantic engine supports lifecycle tracking via ontologies;
- The RegOps engine enables compliance assessment by querying the collected information in a corresponding knowledge graph that captures the AI development and deployment process.

At the seminar, we discussed the relevance of the project's goals to the seminar's core themes, addressing decentralised system governance, ODRL-inspired policy semantics, and verifiable policy modelling in sociotechnical ecosystems.

4.18 The value of Values

Julian Padget (University of Bath, GB) and Harko Verhagen (Stockholm University, SE)

We draw on work in psychology and computer science to propose an approach to the embedding and operationalisation of values – or more precisely, value preference orders – in the design, implementation and operation of STS.

Our motivation is to put forward a methodology – called conscientious design – that puts people at the heart of systems so that (sociotechnical) systems meet – and continue to meet over their lifetime – the expectations of a changing population of participants. A further driver is that for many years we believed that norms were the right technology for capturing and operationalising human requirements in STS, but have concluded that while precise, they are also brittle, hard to write and hard(er) to maintain. In contrast, while not solving the problem, values offer a means to contextualise the norm production and maintenance process to realise what we call small "v" value alignment.

Conscientious design builds upon Schwartz's universal values and Friedman's Value-Sensitive Design (VSD) to propose a frame of reference for STS stakeholder values, in the form of a bespoke value system constructed around the axiology of thoroughness, mindfulness, responsibility. This extends into a process that embeds representations of values that go beyond the design stage to operation, revision and retirement creating a value-based approach to through-life development.

References

1 Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. "Ethical Online AI Systems Through Conscientious Design". In: IEEE Internet Computing 25.6 (2021), pp. 58–64. `https://doi.org/10.1109/MIC.2021.3098324`

2 Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. "Design Heuristics for Ethical Online Institutions". In: Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV. Ed. by Nirav Ajmeri, Andreasa Morris Martin, and Bastin Tony Roy Savarimuthu. Springer, 2022, pp. 213–230. https://10.1007/978-3-031-20845-4_14.

3 Pablo Noriega, Harko Verhagen, Julian A. Padget, and Mark d'Inverno. "Addressing the Value Alignment Problem Through Online Institutions". In: Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI – 27th International Workshop, COINE 2023, London, UK, May 29, 2023, Revised Selected Papers. Ed. by Nicoletta Fornara, Jithin Cheriyan, and Asimina Mertzani. Vol. 14002. Lecture Notes in Computer Science. Springer, 2023, pp. 77–94. https://10.1007/978-3-031-49133-7_5.

4.19 Extending ODRL for AI and Data Regulation

Victor Rodriguez Doncel (Polytechnic University of Madrid, ES)

Elements of the new EU legislation on AI and data can be formalised and operationalised. This opens the door to new types of software tools that support organisations in compliance-related tasks, and also allows for the analysis and simulation of the ethical and societal impacts of emerging technologies and their regulation within STS–this is the goal of the EU HARNESS project. In particular, certain norms can be represented using policy languages such as ODRL. Originally developed as a Rights Expression Language, ODRL has evolved into a more general policy language and could be extended to represent concrete legal norms found in recent AI and data regulations. To achieve this, new language features should enhance ODRL's expressiveness, and the behaviour of ODRL processors should be more precisely defined. Additionally, other ODRL-related tools should be explored, such as: translation mechanisms between ODRL and other languages (e.g., Prolog); methods for efficiently extracting rules from normative texts; and techniques for generating natural language (e.g., English) descriptions from formal rules.

4.20 Compliance Mechanism using LLM

Ken Satoh (Research Organization of Information and Systems, JP)

We launched the tri-lateral (Japan–France–Germany) research project *Research on Realtime Compliance Mechanism for AI (RECOMP)* (https://research.nii.ac.jp/RECOMP/) for the period 2021–2023, supported by the Japan Science and Technology Agency (JST), the Agence nationale de la recherche (ANR), and the Deutsche Forschungsgemeinschaft (DFG).

Our goal is to improve the reliability of AI in society by implementing real-time compliance mechanisms for legal and ethical norms. In our approach, legal norms are modelled as **hard constraints** that must always be satisfied, while ethical norms are modelled as **soft constraints** that should be satisfied as far as possible.

The overall agent architecture is shown in Fig. 1. When a new observation is received, the agent revises its current plan by integrating the new information with a causal theory

that encodes physical constraints and action–effect rules. We then verify legal compliance among candidate plans and select those that are legally valid. Next, we evaluate ethical compliance to identify ethically optimal plans, meaning plans that satisfy stronger ethical rules whenever possible.

In this abstract, I outline a proposal to extend our system using LLMs. At present, our framework is based on logic programming, which creates a gap between the formalised representation and the norms, typically written in natural language. This makes it difficult for both domain experts and system developers to fully understand the underlying logic-based knowledge representation. To address this, we propose the architecture shown in Fig. 2. In this design, system developers provide causal theories and legal and ethical rules in natural language. An LLM then translates them into a corresponding logic program. Similarly, user goals given in natural language are automatically translated into logical goals, and the inferred results from the logic program are translated back into natural language. This approach aims to create a more robust, accessible, and user-friendly system.

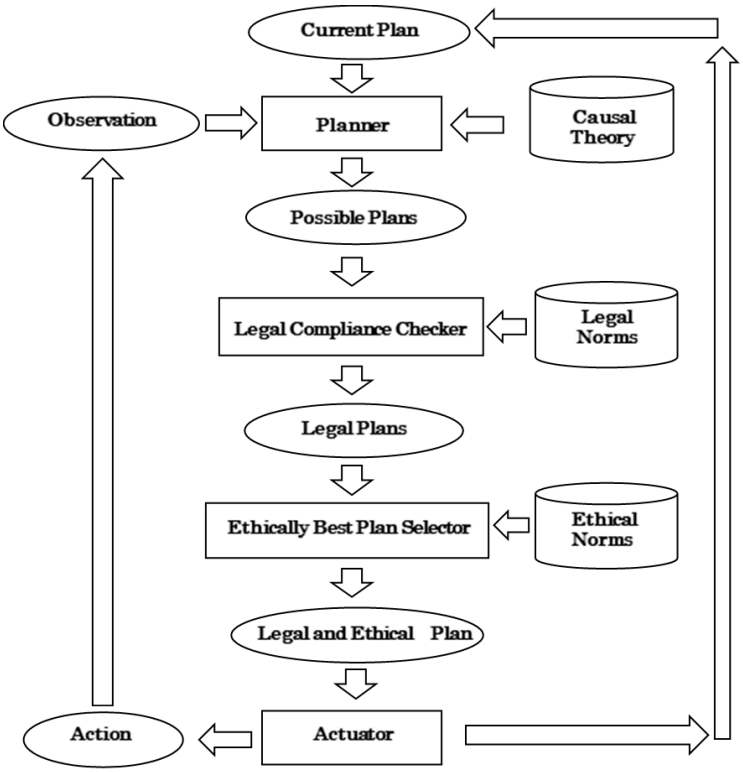

Figure 1 The architecture for RECOMP planning.

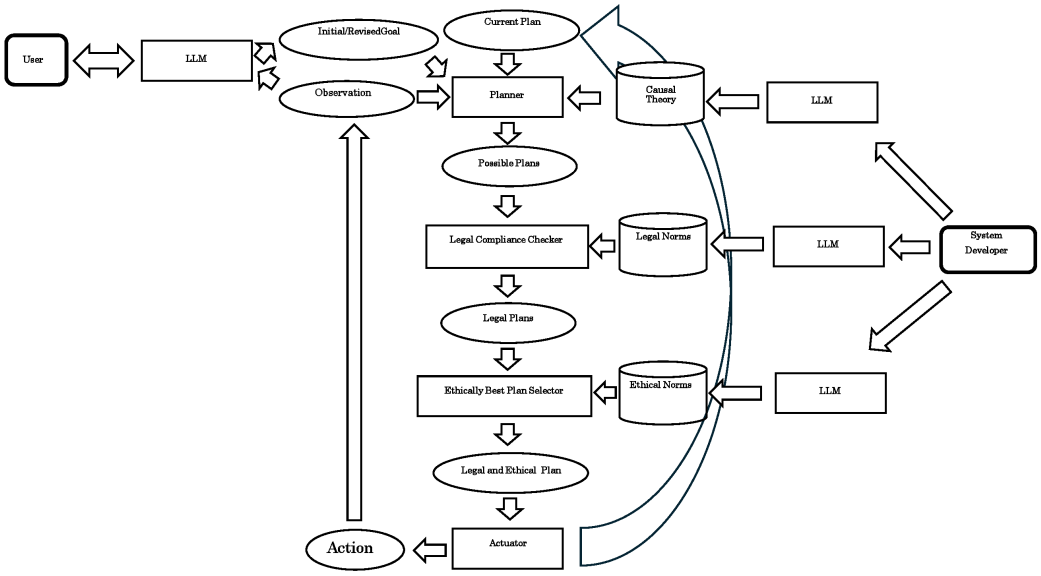

Figure 2 The architecture for RECOMP planning with LLM.

4.21 Regulation in Techno-Human Systems of the Future

Jaime Sichman (University São Paulo, BR)

In our current life, people interact with each other and with several institutions by using technical systems. These complex systems, composed by people and software / hardware are known as STS [1]. As an example, in our University researchers and students access the technical systems to see their grades, submit their works, submit their reports, access the restaurant menu, i.e., to interact with the institution USP.

The current state of the art in AI and Computer Science includes (i) LLMs & Normative Agents, (ii) Big Data for Smart Cities, (iii) Autonomous Systems, (iv) Security & Safety, (v) Green Computing, (vi) Complex Software Supply Chains and (vii) High Performance Computing & Simulation. However, these currently used techniques are not yet integrated in a single framework in order to enable better people's experiences in Digital Society.

Our current work intends to use multi-agent regulation techniques [2] to enhance these STSs. We intend to apply these techniques in the industry 4.0 domain [3] and integrate them with argumentation techniques [4].

References

1 F. E. Emery, E. L. Trist. Socio-technical systems. *Management science, models and techniques*, 2, 83-97, 1960.
2 E. Yan, L. G. Nardin, O. Boissier, J. S. Sichman. A Unified View on Regulation Management in Multi-Agent Systems. *In: Proc. of COINE2025 Workshop*, Detroit, USA, May 2025.
3 Normative Artificial Intelligence for Regulating Manufacturing. `https://naiman.wp.imt.fr/`, Accessed: 01/07/2025.
4 UNBIAS Team – argUmentation and Norm Based Intelligent AgentS. `https://thus.ime.usp.br/teams/unbias`, Accessed: 01/07/2025.

4.22 Ethics and AI: Resisting the Seduction of Frictionlessness

Judith Simon (Universität Hamburg, DE)

1. **Ethics:** Ethics in general asks questions about what is right and wrong, what is good and what is bad, what we should or must (not) do – and for what reasons. If applied to AI, this entails the question of what good and bad AI is – not only technically, but also morally and what we can and/or should (not) use AI for – and for what reasons.
2. **AI:** Looking into the history of AI, we can dissect three themes, related to epistemic, ontological and ethical questions as well as three promises, all of which were assessed later.
   The three themes are:
   - Big data & statistical reasoning: from means and standard deviations to personalisation without subjects
   - The role of imitation and deception
   - The human/cognition/language as both a benchmark and being deficient The three promises of AI are to increase the efficiency, quality and convenience.
3. **Ethical Challenges of AI:** I then outlined some of the most pressing challenges resulting from AI based upon three publications (Deutscher Ethikrat 2023, Simon et al. 2024, Simon 2025).
   These were:
   - Expanding/Reducing Agency
   - AI-based Knowledge Generation & Prediction
   - Endangering the Individual Through Statistical Stratification
   - Effects of AI on Human Competencies and Skills
   - Privacy & Autonomy versus Surveillance & Chilling Effects
   - Data Sovereignty and Data Use Oriented Towards the Common Good
   - Critical Infrastructures, Dependencies and Resilience
   - Path Dependencies & Dual Use
   - Bias and Discrimination
   - Transparency and Accountability – Control and Responsibility
   - Deception

   Having outlined the most pressing ethical challenges of AI, I argue that the judgment on the increased quality of cognitive process and decision making is still open and differs for different individuals & groups. The judgment on whether (Gen)AI increases the efficiency & convenience is also open – but even if epistemic processes were more convent and efficient, this very improvement comes with epistemic, ethical and political costs.
4. **Conclusions:** I concluded my talk with some suggestions on what we can do to design technologies with ethics in mind – but while also being aware about the limits of reaching ethical and political goals with and through technology.
   These are the following:
   - There is no "machine ethics": Ethics can't be delegated to machines, but requires judgment, situated and context-aware reasoning.
   - Another way of thinking about ethics and AI is rooted in the "Values in Design" approach. Instead of delegating ethics to tech, it asks: which values are relevant for whom and how can and should we operationalise them?

- Ethics is not a check-box to tick. Instead ethical considerations are part and parcel of the whole life-cycle of developing a deploying tech: from creating and annotating data, to choosing methods, using tech in specific contexts and taking care of their remains after they cease to work.
- So ethics is part of research and tech development, but also goes beyond tech. Think of de-biasing AI – you can't fully avoid discriminating against every possible group or individual, but have to make ethical and political choices, which harms are most important to prevent.
- Finally: Beware of the pitfalls of anthropomorphising technology and making humans machinic. Both are inherent in the history of AI anyway, but become even more salient in the field of normative MAS and ethical AI.

References

1 Deutscher Ethikrat (2023). *Mensch und Maschine – Herausfoderungen durch Künstliche Intelligenz*. Available at: `https://www.ethikrat.org/en/publications/opinions/humans-and-machines/`
2 Simon, J., Spiecker gen. Döhmann, I. & von Luxburg, U. Generative KI – Beyond Dystopia and Simple Solutions. Discussion Paper No. 34. Halle (Saale): National Academy of Sciences, Leopoldina, 2024. doi:10.26164/leopoldina_03_01245
3 Simon, J. (2025). Generative AI, Quadruple Deception and Trust. Social Epistemology. doi: 10.1080/02691728.2025.2491087
4 Simon, J., et al. (2020). Algorithmic bias and the Value-Sensitive Design approach. Internet Policy Review, 9(4), 1–16. Available at: `https://policyreview.info/concepts/algorithmic-bias`

4.23 Flexible, adaptive sociotechnical systems based on norms and values, their evolution, and their realisation using generative AI

Munindar P. Singh (North Carolina State University – Raleigh, US)

Our computational model of STS enables a natural way to elicit stakeholder needs (requirements, risk attitudes, and value preferences), reason about them, and build decentralised MAS meeting those needs. We have been working on improvements of this model to incorporate enhanced reasoning about value preferences and norms, especially in conjunction with each other and with mental constructs such as goals. In this seminar, we will discuss ideas relating to (1) a lifecycle for STS, especially its continual tracking and alignment with potentially changing stakeholder needs, (2) models for the emergence and evolution of norms in light of both observations and semantically rich models, and (3) realising STS by taking advantage of the facilitation of knowledge engineering provided by generative AI, including identifying ways to enrich current generative AI models and toolkits with social intelligence, thereby achieving a leap in the development of MAS that go far beyond today's rigid, workflow-based approaches.

4.24 Architects of Trust: A Framework for Sovereign Data Governance in the Age of Autonomous Agents

Simon Steyskal (Siemens AG – Wien, AT)

The emergence of decentralised STS, from industrial data spaces to autonomous MAS powered by LLMs, has exposed fundamental limitations in traditional, centralised governance models. This work presents a conceptual framework addressing the critical research challenge of establishing trustworthy, policy-based governance in environments where autonomous agents must interact without central authority.

The framework synthesises two W3C standards in a novel "two-pillar" architecture: the Open Digital Rights Language (ODRL) for expressing deontic policy semantics, and the Shapes Constraint Language (SHACL), repositioned as a dynamic policy enforcement engine through custom node expressions. We outline how this integration, combined with Decentralised Identifiers (DIDs) and Verifiable Credentials (VCs), could enable verifiable, context-aware governance that preserves data sovereignty while facilitating automated compliance checking.

Key research directions identified include: (1) developing formal semantics for ODRL-aware SHACL validation that transforms static data validators into Policy Decision Points, (2) addressing semantic gaps in current policy languages for complex temporal and contextual constraints, (3) establishing mechanisms for policy conflict resolution in multi-stakeholder environments, and (4) extending governance models to encompass LLM-powered agents where policies themselves may be generated through natural language interaction.

4.25 Human-Centred AI in Sociotechnical Systems

Sz-Ting Tzeng (University of Umeå, SE)

As AI increasingly integrates into our social structures, humans and AI form complex STS. Ensuring that AI aligns with human values and social norms and that AI behaviours are justifiable becomes increasingly important when humans are involved. My research focuses on developing human-centred AI that can make decisions and adapt its explanation strategies according to the social context and values in STS. In this seminar, I explore how decision making and AI-generated explanations reflect and are shaped by human values, and how agents adapt to evolving STS.

4.26 Ethical Decision-Making in Multi-Agent Systems

Jessica Woodgate (University of Bristol, GB)

Consequential decision-making is increasingly guided by AI in diverse social settings, from resource allocation to balancing preferences of stakeholders. Whilst AI has beneficial uses,

its sociotechnical nature entails it often adopts default social norms (standards of expected behaviour) and power structures of society, which includes systematic injustices and inequalities. Resource allocation may treat some recipients more favourably, or the preferences of minorities may be overlooked. Realising the benefits of AI across society necessitates addressing ethical implications, understood as what is morally good or right. Many ethical concerns are multi-agent in nature, involving one party's concern for another. MAS, which are collections of multiple agents interacting in a shared environment, are thus an appropriate setting to examine ethical implications of AI and encompass social factors such as norms. To advance ethical decision-making in MAS, operationalising principles from normative ethics – the philosophical study of practical means to determine right from wrong – helps support interdisciplinary insights and guide decision-makers in making evaluative judgements.

4.27 Sociotechnical reasoning of privacy

Pinar Yolum (Utrecht University, NL)

STS consist of agents and humans, each with potentially different capabilities, working together to accomplish tasks. For these systems to succeed, agents need to recognise, take into account, and demonstrate social, developmental, and communication skills – like self-reflection and empathy – that are typically linked to humans. How do we realise STS that benefit from these skills? How do we measure their existence? I argue that our vocabulary for talking about STS is based on individual AI systems and do not capture the effect of such skills. I demonstrate a few cases over the domain of privacy.

4.28 Regulating autonomous agents on the Web

Antoine Zimmermann (Ecole des Mines – St. Etienne, FR)

Much digital activity happens on the Web. As the Web gets increasingly vast and complicated, we need assistance from automated tools to process the information and sometimes do tasks for us. Complicated tasks require proactiveness and autonomy to complete automatically. When autonomous software acts on our behalf on the Web, it must do so in accordance with our policies and regulation. Therefore, we need mechanisms that allow artificial agents to become aware of their obligations, permissions, and prohibition in very strictly verifiable ways. Language models can help translating human-readable regulation into machine-processable representations, but they are prone to misinterpretation, approximation, and hallucination. We want to convey policies to agents on the Web in a formal, unambiguous form, and make them easily discoverable in a systematic way, such that agents can arrive at a location on the Web and operate on available resources according to the rules, without prior knowledge of the local context.

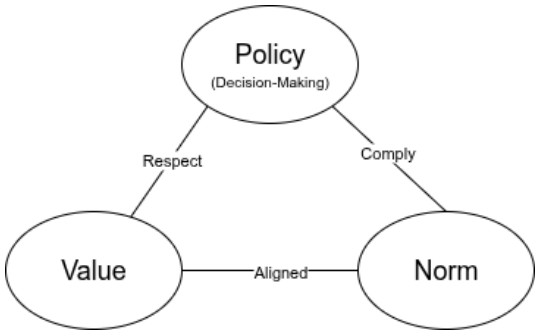

Figure 3 General interplay between policies, values, and norms.

5 Working groups

5.1 Coevolution of Values and Norms in Sociotechnical Systems

Nirav Ajmeri (University of Bristol, GB), Marina De Vos (University of Bath, GB), Davide Dell'Anna (Utrecht University, NL), Pradeep Murukannaiah (TU Delft, NL), Vivek Nallur (University College Dublin, IE), Luis Gustavo Nardin (IMT Mines Saint-Étienne, FR), and Munindar P. Singh (North Carolina State University – Raleigh, US)

5.1.1 Introduction

In this report, we provide an overview of the discussions held at the Dagstuhl 25721 Working Group on the coevolution of values and norms in a STS. We discuss the background and conceptualise an STS where norms and values can change over time and influence each other based on the observations and actions of the human actors and agents. We identify key research challenges spanning the formalisation, operationalisation, and application of such an STS.

An STS involves social actors (*humans*) and technical entities (abstracted as AI *agents*) [29]. The agents represent the social actors and aim to facilitate rich interactions among them. Two key factors that influence social actors' interactions in an STS are **values** and **norms**. Values represent deep-rooted motivations or preferences of social actors (to act in a certain way). In contrast, norms govern expectations between actors. Norms usually reflect the values of the social actors, but they can also shape the values of social actors. Both values and norms influence the policies agents adopt for decision making as shown in Figure 3.

Literature on engineering STS studies the evolution of values and norms independently, e.g., [18, 22]. However, their interplay (bidirectional) – how values inform norms and how norms influence values [30] – is largely unexplored. In this report, we identify key research avenues to conceptualise this interplay, model an STS with this conceptualisation, and engineer the agents in the STS to co-evolve values and norms.

5.1.2 Background

Values are generally considered as high-level motivations that drive human behaviour [33]. Value preferences describe the relative importance that a human ascribes to different values to guide their actions in a socio-cultural environment and context. Values in society are operationalised at the agent-level by aligning their actions with individuals' value preferences, and at the STS level by expressing and enforcing norms to regulate agents' behaviour and interactions [6, 19].

Research in (normative) MAS has explored several approaches to model and compute the norms required to make coordination between agents possible [8], to address norm violation and sanctioning [17, 31, 2, 1, 36], and to support aspects pertaining to the dynamic adaptation of norms, including a variety of centralised and distributed approaches for norm change, revision, emergence, and learning [4, 10, 32, 11, 5, 16, 13, 27, 24, 39, 38].

Research has also explored approaches to infer human values [22, 23] and to relate norms to values [34, 20, 37, 3]. Further, values have been shown to affect policy and norms [12]. Despite the extensive literature on norms and values, the study of the interplay and co-evolution of norms and values over time still remains largely under-explored.

5.1.3 Main discussion points

This section summarises the key points that emerged from our discussion. This discussion led us to the conceptualisation and research questions outlined later.

1. Distinction between evolution, adaptation, emergence, learning, and change of norms.
   - Evolution is emergent, and adaptation is more top-down.
   - Evolution is a gradual process, while adaptation does not have to be.
   - Adaptation refers to a change in norms.
2. Distinction between top-down and bottom-up norms creation.
   - In the top-down case, norms come from an institution that regulates agents' behaviour. In the bottom-up case, norms emerge or are agreed upon by the agents themselves in an agent-centric (distributed) regulation.
3. What triggers the need for adaptation?
   - Norms may change when they become incompatible with values. Values may change based on the perception of others.
4. A norm is understood to have emerged when a certain percentage of the population (e.g., 90%) adopts it [21]. This notion refers to the idea of social tipping points.
5. The size of inner and outer groups affects the strengths of sanctions, the vulnerability perception of inner groups, and the willingness to violate norms.
6. There is a difference between revealing actions and values underlying those actions.
7. While there is some work on norm change ([28, 25, 14, 7, 26]), less work is present on value change.
8. The co-evolution of norms and values is relevant for a variety of case studies. These include
   - Hospital and Healthcare scenarios [9], where evolving human norms and values need to be considered to ensure that technology adequately supports patients and citizens. As an example, we considered *Family bounding* and *privacy* as possible values in this context. Norms could relate to permissions to share data with family. Preferences express individuals' preferences over values, such as *family bounding* over *privacy*.

- Human-AI teams [35, 15], where it is essential that team members have a common understanding of other team members and their individual preferences and values, team norms and team objectives (incl. team and organisational values) to ensure that the team operates (increasingly) effectively over time and that both AI agents can adequately support humans in the teams.
- University settings, where individual norms and values interact and coexist with organisational values.
- Legal settings, where personal norms and values co-evolve with societal values and legal norms. We consider a scenario where an individual values *traditional family*, and legal norms in their country prohibit abortion. Individual actions could include abortion (an action that violates a norm within a traditional family), but also protest against abortion (an action that is in line with their values, but inconsistent with other actions). One such situation expresses a choice to violate the norm without the desire to change it from an individual point of view. This situation could be an indication of an incoherence between individual values and norms, potentially leading to a change in one or the other.

9. Considering multiple actions being executed by multiple agents concurrently or sequentially (i.e., the temporal aspect of action execution from multiple agents) is important to highlight the sociotechnical nature of the problem of norm-values coevolution and their relation with agents actions, and the multiplicity and interactions between agents. As a simple illustrative example, we considered the case of a fight resulting from two persons moving their hands at the same time.
10. If we were to integrate norms and values in an agent architecture, such as a BDI agent, where would they live? According to Schwartz [33], values can be interpreted as beliefs. The norms literature sometimes associates values with Desires, although implementations may not follow this. Most of the literature considers Beliefs as information, knowledge, while Desires are considered objectives to achieve/comply with if possible, Intentions are intended as plans to follow and adhere to.
11. Norm changes can lead to values change. Popularity is not enough to change norms, the general change in values may lead to the change in norms. Accountability is essential to a norm. Values may not even change for attitudes to change. What could change is how human/agent sees or *perceives* the values in the view of new experiences or situations.
12. Affordance may enable the change of attitudes but also may make possible or not change of values and norms.

5.1.4 Conceptualisation

Figure 4 shows our conceptualisation of how values and norms interplay in an STS. To start with, each agent is endowed with a representation of the values of its principal. This endowment can happen in various ways, including learning from the human-agent interaction. Similarly, the STS also starts with a set of norms specified by the organisation. These initial norms can be established or learned via, e.g., the negotiations among the stakeholders.

In this setting, one key challenge is to enable an agent to acquire a policy (how to act) that aligns with its principal's values and complies with the organisational norms. However, often a policy may not be able to accomplish both objectives, i.e., (sequences of) action(s) may align with a value but deviate from a norm, or comply with a norm but deviate from values. These circumstances provide an opportunity for value and/or norm revision.

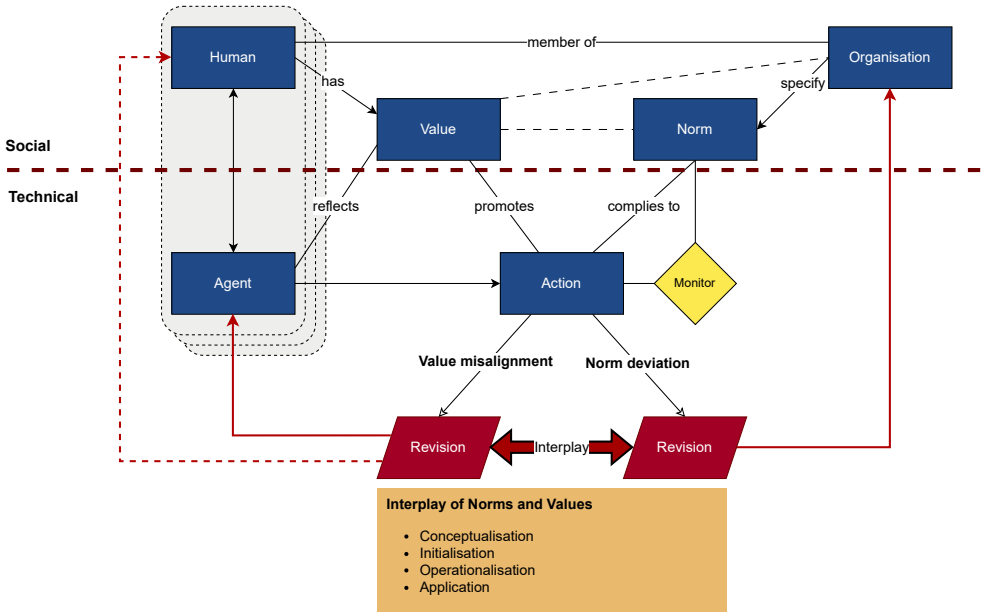

Figure 4 A conceptualisation of the interplay between policies, values, and norms in an STS.

We understand norm revision as a process that can be executed by an organisation in the STS that specifies the norms. Norm-revision can be informed by a variety of aspects, such as the values of individual agents in the organisation (if known), by the observation of agents complying or violating the norms, by their effectiveness in achieving organisational objectives, and by aspects such as popularity or affordances. Value revision, on the other hand, happens internally to the individual agents and can take different forms, such as changing the preference order between different values or changing (increasing or decreasing) the strength of a preference. Value revision, in this sense, is informed by the current norms that are enforced in the STS, by the human the agent is representing, and by the actions available to the agent, among others.

5.1.5 Challenges

Norms and Values alignment and interplay

This category of challenges refers to the conceptualisation of the problem of norm and value change and their interplay.

Human/Conceptual/Theoretical aspects

- What is the link between values and norms? We consider interplay in both directions, i.e., norms to value, and values to norms?
- How do affordances affect values?
- How does value change and clarification drives norm change?

Computational aspects

- How to model and reason about the alignment between agent policies, norms, and values in Intelligent Agents?
- How to characterise the interplay between norms and values when different types of norms (e.g., social vs more regulative norms) are considered?

- How to measure (in)coherence between an agent's actions, norms and values? For example, the temporal aspect matters when evaluating actions with respect to values.

Computational triggers and reasons for change

This category of challenges refers to the modelling of computational triggers that may cause norm and value change.

- When (if) to change norms and values?
- To what extent do public/private norm violations or compliance can initiate change?
- What are possible triggers/drivers of policy/value change?
- Observation of violation, compliance, sanction, other agent's actions
- What is a mechanism that uses a measure of incoherence/misalignment/asymmetry between norms and values as a pressure for change?

Computational norms and values change

This category of challenges refers to the computational mechanisms to implement actual norms and values change.

- How to change norms and values (value preferences rather than values)?
- How to ensure norms change but still within the boundaries and objectives of the intended system?
- How to ensure the system remains fit for purpose?

Computational dynamics and effects of change, from local to system-level and back

This category of challenges refers to the effects of norms and values change on the agents and on the STS as a whole, and to the resulting dynamics between norms, values and agents behaviour. These dynamics could be studied in a controlled setting for instance via (social) simulations, and in less controlled settings via longitudinal human-AI interaction studies.

- How can change in norms and values propagate from local groups to the larger society/organisation, and possibly back?
- When/how to move from convention to social norms?
- How to facilitate "integration" of new members (with their norms and values) into a group, and how new members affect the group norms and values?
- Can an AI that is able to reason about alignment of actions to norms and values and their dynamics, help people expose their reasoning about their values?

5.1.6 Future Plans

Our next steps involve (1) refining the research challenges identified above into a structured research agenda, and submitting it to, for instance, the AAMAS Bluesky track; (2) forming smaller working groups to focus on specific research areas; (3) organizing an online seminar series; and (4) preparing research proposals for funding calls.

References

1 Rishabh Agrawal, Nirav Ajmeri, and Munindar P. Singh. Socially intelligent genetic agents for the emergence of explicit norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 10–14, Vienna, July 2022. IJCAI.

2 Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 28–34, Stockholm, July 2018. IJCAI. doi:10.24963/ijcai.2018/4.

3 Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 16–24, Auckland, May 2020. IFAAMAS. doi:10.5555/3398761.3398769.

4 Giulia Andrighetto, Sergey Gavrilets, Michele Gelfand, Ruth Mace, and Eva Vriens. Social norm change: drivers and consequences, 2024.

5 Duangtida Athakravi, Domenico Corapi, Alessandra Russo, Marina De Vos, Julian Padget, and Ken Satoh. Handling change in normative specifications. In *International Workshop on Declarative Agent Languages and Technologies*, pages 1–19. Springer, 2012.

6 Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

7 Jordi Campos, Maite Lopez-Sanchez, Maria Salamó, Pedro Avila, and Juan A. Rodríguez-Aguilar. Robust Regulation Adaptation in Multi-Agent Systems. *ACM Transactions on Autonomous and Adaptive Systems*, 8(3):1–27, September 2013. doi:10.1145/2517328.

8 Amit Chopra, Leendert van der Torre, Harko Verhagen, and Serena Villata. *Handbook of normative multiagent systems*. College Publications, 2018.

9 Amit K Chopra and Munindar P Singh. Accountability as a foundation for requirements in sociotechnical systems. *IEEE Internet Computing*, 25(6):33–41, 2021.

10 Rosaria Conte, Giulia Andrighetto, and Marco Campennì, editors. *Minding Norms: Mechanisms and dynamics of social order in agent societies*. Oxford University Press, 2013.

11 Domenico Corapi, Alessandra Russo, Marina De Vos, Julian Padget, and Ken Satoh. Normative design using inductive learning. *Theory and Practice of Logic Programming*, 11(4-5):783–799, 2011.

12 Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: values, norms and culture in multi-agent systems. *Artificial intelligence and law*, 21:79–107, 2013.

13 Davide Dell'Anna, Natasha Alechina, Fabiano Dalpiaz, Mehdi Dastani, and Brian Logan. Data-driven revision of conditional norms in multi-agent systems. *Journal of Artificial Intelligence Research*, 75:1549–1593, 2022.

14 Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. Runtime revision of norms and sanctions based on agent preferences. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 1609–1617. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

15 Davide Dell'Anna, Pradeep K Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M Jonker, Catharine Oertel, and Pınar Yolum. Toward a quality model for hybrid intelligence teams. In *23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024*, pages 434–443. ACM Press Digital Library, 2024.

16 Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. Runtime revision of sanctions in normative multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 34:1–54, 2020.

17 Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. Is this a violation? learning and understanding norm violations in online communities. *Artificial Intelligence*, 327:104058, 2024.

18 Paul R. Ehrlich and Simon A. Levin. The evolution of norms. *PLOS Biology*, 3(6), 2005. doi:10.1371/journal.pbio.0030194.

19 Sven Ove Hansson. *The structure of values and norms*. Cambridge university press, 2001.

20 Samaneh Heidari, Maarten Jensen, and Frank Dignum. Simulations with values. In *Advances in Social Simulation: Looking in the Mirror*, pages 201–215. Springer, 2020.

21 James E Kittock. Emergent conventions and the structure of multi-agent systems. In *Proceedings of the 1993 Santa Fe Institute Complex Systems Summer School*, volume 6, pages 1–14. Citeseer, 1993.

22 Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. Value inference in sociotechnical systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1774–1780, London, 2023. IFAAMAS.

23 Enrico Liscio, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Value preferences estimation and disambiguation in hybrid participatory systems. *Journal of Artificial Intelligence Research*, 82, April 2025. doi:10.1613/jair.1.14958.

24 Mehdi Mashayekhi, Nirav Ajmeri, George F. List, and Munindar P. Singh. Prosocial norm emergence in multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 17(1–2):3:1–3:24, June 2022. doi:10.1145/3540202.

25 Mehdi Mashayekhi, Hongying Du, George F. List, and Munindar P. Singh. Silk: A simulation study of regulating open normative multiagent systems. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 373–379. AAAI Press, 2016.

26 Javier Morales, Michael Wooldridge, Juan A. Rodríguez-Aguilar, and Maite López-Sánchez. Off-line synthesis of evolutionarily stable normative systems. *Autonomous Agents and Multi-Agent Systems*, June 2018. doi:10.1007/s10458-018-9390-3.

27 Andreasa Morris-Martin, Marina De Vos, and Julian Padget. Norm emergence in multiagent systems: A viewpoint paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 33(6):706–749, 2019.

28 Andreasa Morris-Martin, Marina De Vos, Julian Padget, and Oliver Ray. Agent-directed runtime norm synthesis. In *AAMAS23*, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

29 Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1706–1710, Auckland, 2020. IFAAMAS.

30 Pradeep K. Murukannaiah and Munindar P. Singh. From machine ethics to internet ethics: Broadening the horizon. *IEEE Internet Computing*, 24(3):51–57, 2020. doi:10.1109/MIC.2020.2989935.

31 Luis G Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K Kalia, Jaime S Sichman, and Munindar P Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review*, 31(2):142–166, 2016.

32 Bastin Tony Roy Savarimuthu and Stephen Cranefield. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems*, 7(1):21–54, 2011.

33 Shalom H Schwartz and Wolfgang Bilsky. Toward a universal psychological structure of human values. *Journal of personality and social psychology*, 53(3):550, 1987.

34 Marc Serramia, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. A qualitative approach to composing value-aligned norm systems. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems*, pages 1233–1241, 2020.

35 Munindar P Singh. The intentions of teams: Team structure, endodeixis, and exodeixis. In *ECAI*, volume 98, page 303. Citeseer, 1998.

36 Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. Norm enforcement with a soft touch: Faster emergence, happier agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1837–1846, Auckland, May 2024. IFAAMAS. doi:10.5555/3635637.3663046.

37 Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. Value-based rationales improve social experience: A multiagent simulation study. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, pages 3612–3619, Santiago de Compostela, October 2024. IOS Press. doi:10.3233/FAIA240917.

38 Jessica Woodgate and Nirav Ajmeri. Combining normative ethics principles to learn prosocial behaviour. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1–3, Detroit, May 2025. IFAAMAS.

39 Jessica Woodgate, Paul Marshall, and Nirav Ajmeri. Operationalising Rawlsian ethics for fairness in norm-learning agents. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2382–26390, Philadelphia, February 2025. AAAI. doi:10.1609/aaai.v39i25.34837.

5.2 Social Agentic Systems

Matthew Arrott (Coactive Computing, US), Matteo Baldoni (University of Turin, IT), Victor Charpenay (Mines Saint-Étienne, FR), Amit K. Chopra (Lancaster University, GB), Timotheus Kampik (SAP Berlin, DE & Umeå University, SE), Nadin Kokciyan (University of Edinburgh, GB), Ken Satoh (Research Organization of Information and Systems – Tokyo, JP), Jaime Sichman (University São Paulo, BR), Munindar P. Singh (North Carolina State University – Raleigh, US), and Pinar Yolum (Utrecht University, NL)

5.2.1 Introduction

Agentic computing is a recent paradigm for leveraging generative AI to build agents that accomplish sophisticated tasks. Today's agentic systems are impoverished because of a focus on rigid coordination methods based on procedural encodings of interaction, such as via task graphs. We introduce the notion of social agentic systems (SASY). SASYs are systems of agents that explicitly represent and communicate about norms, including the consequences of respecting or violating a norm. When needed, agents may deviate from norms and can negotiate their relaxations. In this way, SASYs provide a way to leverage the flexibility and power of generative AI through high-level, socially inspired models of interaction.

With the advent of software systems based on generative AI and LLMs, autonomous agents and MAS have re-emerged as one of the key research fields that will influence the future of large-scale information systems. Indeed, substantial expectations are attached to the importance that AI agents will have on how individuals and organisations carry out collaborative knowledge work. At the same time, concerns are growing that unjustified expectations regarding mainstream agent technologies will lead to failures of agent introduction projects; relatedly, the labelling of conventional (i.e., non-agent) technologies as "agents" is sometimes called *agent washing* [12].

These concerns raise the question of what is fundamentally needed to deploy software agents successfully as part of STS [17], such that human productivity and well-being are

indeed affected substantially and sustainably. In this chapter, we argue that a key capability – missing in current mainstream agent and MAS architectures and frameworks – is the ability to represent, reason, and communicate about norms and social values, in order to effectively operate and evolve the STS. We call STS that feature software agents with such capabilities social-agentic systems (SASY).

Based on a motivating example, we explain why the aforementioned capabilities are crucial for the successful deployment of intelligent software agents within and across organisations. We then explain why and how SASY differ architecturally from classical MAS, as well as from LLM-based agent architectures. Finally, we outline a set of research challenges that can move us towards real-world SASY.

5.2.2 Agentic Systems

An LLM-based chatbot is not only able to interact with humans, but it can also interact with technical systems by generating structured output such as code snippets: the chatbot generates a code snippet that is executed in a sandbox environment, giving controlled access to the technical system, and the result of the execution is given back to the chatbot in a textual form. In this setting, the chatbot becomes an agent, perceiving and acting on its environment [19].

Many agent frameworks have been built on this idea in the past two years, including LangGraph, AG2 (previously branded AutoGen), the OpenAI Agents SDK, Smolagents, and CrewAI. These frameworks encourage modularity in diverse respects. An agent interacts with its environment via a collection of *tools*, each providing access to a particular functionality. For example, LangGraph provides built-in tools for Web search, Web scraping, API access, code interpretation, and database access [11]. Likewise, the Model Context Protocol (MCP) [1] facilitates the integration of external tools into an agent's environment. Another way modularity is encouraged is by decomposing task handling into components, each component being made of a chatbot with its own context. Such a modular agentic system is sometimes referred to as a "MAS" in this literature (though at variance with the terminology in the MAS community), though from a more general point of view, it is indistinguishable from a single agent.

Even though LLM-based agentic systems perform much more poorly than humans, they achieve surprisingly good performance on several benchmarks. For instance, on the General AI Agent (GAIA) benchmark, GPT-4 correctly answers 30% of the questions (whose answers require Web search and reading documents in various formats) [14]. On the more general AgentBench benchmark, GPT-4 and Claude 3 achieve 14% to 70% of the tasks, depending on the domain, ranging from puzzle solving to Web browsing [13]. OpenAI and Anthropic provide commercial support for personal assistants evolving in a Web or computer environment. Planning a trip, which includes online reservation and payment, is one of the use cases advertised by the two companies, for example. Such a use case may – or, rather, must – include numerous social interactions (via the Web). Yet, LLM-based agentic systems demonstrate no form of social awareness.

5.2.3 Motivating Example

Let us consider the following scenario:

> A researcher Jaime, who works at the University of São Paulo, located in Brazil, is invited to a Dagstuhl Seminar. Pınar, a researcher who works at Utrecht University, located in the Netherlands, is invited to the same seminar. The two of them are preparing a joint research proposal, and hence have agreed to have a first meeting on the Sunday evening prior to the seminar.

To promote *sustainability*, Utrecht has an internal regulation that trips to cities that are less than 700 km away from Utrecht must be made by train. To promote *reasonable usage of public resources*, São Paulo has an internal regulation that researchers must buy economy flights. Additionally, Jaime has a preference to always fly with a certain airline, since his membership in its loyalty program allows him to upgrade his ticket. Both Jaime and Pınar would like to book a taxi together from the train station at Türkismühle to Dagstuhl.

In principle, this scenario presents *social constructs* that have been studied by the MAS community for the last 40 years [10, 9, 7]:

Institutional norms that must be taken into account during the agents' deliberation and decision-making.

Institutional or individual values to consider in choosing a solution.

Commitments between two parties, representing that one party legitimately expects that the counterparty will act accordingly.

Individual preferences to prioritise when several solution options are feasible.

Suppose the train that Pınar has booked is late, and this would prevent her from sharing a taxi with Jaime, as they initially planned. Suppose also that the taxi service in Türkismühle closes at 20:00 on Sundays and reopens on Monday morning. Current agentic systems do not appropriately cope with such diverse social concepts and changes in plans.

An intelligent agent would need to reconsider, either autonomously or by asking the user, which adaptations should be made to the original joint plan: should Pınar take an earlier train? Should she take a flight instead of a train? That is, how might she deliberate on whether to violate a norm to guarantee the meeting? Moreover, if a taxi reservation has been made for an earlier time, this commitment should be revised and communicated to the taxi service.

We posit that leveraging such *social constructs* is crucial to addressing the challenges facing agentic AI.

5.2.4 SASY Architecture

We define a SASY as a MAS consisting of human and software agents, in which software agents engage in communication dialogues that run over extensive periods of time and in which their actions, including speech acts, are grounded in norms, values, and trust (or lack thereof). The internal structure of agents in SASY could be designed in various ways; that is, a SASY may feature agents of diverse architectures. We assume that each agent has access to the environment where it can communicate with other agents. Agents have knowledge of the domain they are operating in, they have capabilities, and can reason about what action to perform. The domain reasoner could be a traditional reasoner, such as a rule-based engine, or a more sophisticated one using the capabilities of neural models. Some agents may have a social state component where they have an explicit representation of social preferences,

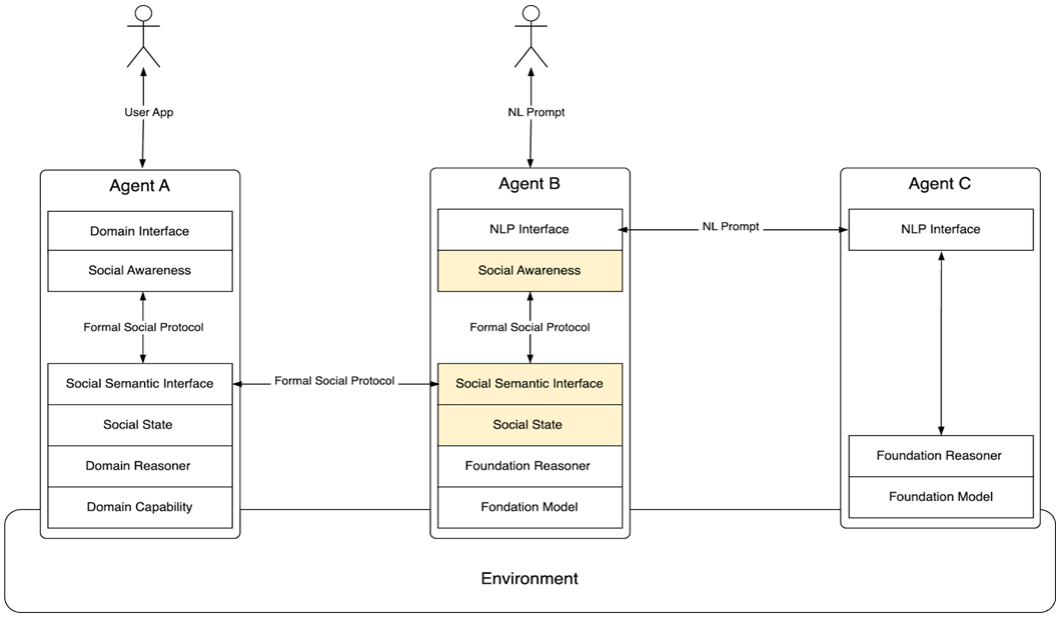

Figure 5 Communication between different types of agents. SASY enables the creation of socially-aware LLM agents, that is, of the type of Agent B.

norms, values, commitments, and so on. An agent may have an interface to interact with users. Such an interface could be a built-in application interface or an interactive interface, such as a chatbot, where the interactions occur through natural language. Moreover, when agents interact with humans, they gather additional information about these humans to become socially aware by encoding the information into formal specifications that could be used to communicate with other agents.

Let us assume we have three different types of agents as depicted in Figure 5. All agents have access to the environment, where they can communicate with each other through well-defined protocols. Agent A is an agent that is using a user application to interact with the user to process the user requests. The *Social Awareness* component includes information such as user preferences. A formal social protocol ensures that the user state is translated to a formal state, which could be used to communicate with other agents in the STS. Agent B supports a natural language interface for the user. Agents B and C conform to the SASY system architecture, since they are both equipped with a natural language interface. Agent B is using a natural language interface to communicate with the user, whereas Agent C processes natural language prompts, but it does not interact with the user directly.

The literature shows examples of Agent A [4] and Agent C [11]; however, research challenges remain. Architectural components that require further attention from the research community are highlighted in Agent B. For example, if Jaime and Pınar both had agents of type Agent B, their agents could communicate with each other and also with their users to adjust the plan according to the current social context.

In the example from Section 5.2.3, a SASY software agent could address the situation as follows:

- Decide that booking a flight is undesirable, not only because it violates the *sustainability* norm, but also because it is, from a commonsense perspective, not *comfortable* in the current situation and not a *reasonable use of public resources*. Whereas the latter two

norms are not formally represented in the norm base the agent has access to, the agent generates them on the fly using an LLM and presents them as explanations for the final proposal to Jaime and Pınar.
- Decide that the revised taxi trip is only marginally noncompliant with the operating hours of Taxi Martin and use this inference as the basis for a successful request (carried out via an interface to the traditional telephone system) that re-negotiates the schedule and conditions for the taxi trip.

5.2.5 Challenges

In order to realise the vision outlined above, agents must exhibit social awareness in human-agent interactions, which raises several research questions:

Prompting. How do we adapt the prompts with social constructs, such as norms and values? What are the different ways to adapt the prompts to ensure social components are represented adequately? Some options are enhancing the prompts, revising the prompts, and so on.

Social interactions. How do we design components to keep track of or regulate social interactions? Can an LLM figure out that it is making a commitment when it is saying certain things to the user? Consider the recent Air Canada case [8], where the airline's chatbot told a customer they could apply for a refund, in contradiction to the company's policy. Could we keep such commitments in a database and keep track of them so that the LLM can decide whether to make or break these commitments with its interactions? Commitments here are one such abstraction; we can also think of consent [2, 18] and other such abstractions similarly. For example, a consent store can prohibit an LLM from generating certain results or communicating what may have been generated.

Social. How may LLMs interact with this social state? Some possibilities are below.
- The LLM generates what it will normally generate and then sends it to the social awareness component, which then checks whether this is appropriate or not. It could be a binary decision or a modification to the output.
- Social awareness component generates (additional) prompts to the LLM to take into account so that the generated output is socially appropriate.

Validation. How do we assess that the proposed architecture delivers intended actions or outputs? Typical LLM evaluation is with benchmarks on input and output. This could be one way to evaluate SASY, but to specifically evaluate the social awareness component or the interface, we might need different techniques. For example, if the SASY produces socially appropriate output, is it because the social awareness component caught something and fixed it? Or, did the LLM generate it that way to begin with?

Languages. What are some languages or protocols that could help in realising this architecture? Formal languages to represent norms [5, 6]remain useful for verifiability and accountability reasons, even in the presence of a natural language interface. The Blindingly Simple Protocol Language (BSPL) [15, 16] provides an alternative to workflows. Translating natural language into a formal language and back is an important challenge. Some LLM services are already capable of outputting structured data, validating a schema defined at run time (e.g., ChatGPT Structured Outputs), but the support is inadequate. A typical problem arising in this context is terminological alignment [3].

In addition to serving as a personal assistant as described above, the agentic architecture could also help in improving various design stages of STS. For example, SASY could be used to simulate various stakeholders of an STS, adopting various personas, depicting

different interactions, and realising diverse scenarios. SASY could be used to simulate various alternative evolutions of an STS. Through such simulations, it may be possible to infer various side effects on primary stakeholders. Moreover, it may help in identifying secondary stakeholders, that is, those who would be affected by the use of the system.

References

1 Anthropic. Model Context Protocol (MCP), July 2025. Accessed 2025-07-11. URL: https://modelcontextprotocol.io/.

2 Anastasia Sophia Apeiron, Davide Dell'Anna, Pradeep K. Murukannaiah, and Pınar Yolum. Model and mechanisms of consent for responsible autonomy. In *Proceedings of the 24th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 133–141, Detroit, May 2025. IFAAMAS. doi:10.5555/3709347.3743525.

3 Victor Charpenay. Génération et validation de données structurées. In *36es journées francophones d'Ingénierie des Connaissances (IC)*, 07 2025. URL: https://hal-emse.ccsd.cnrs.fr/emse-05163532v1/document.

4 Amit K. Chopra, Matteo Baldoni, Samuel H. Christie V, and Munindar P. Singh. Azorus: Commitments over protocols for BDI agents. In *Proceedings of the 24th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 490–499, Detroit, May 2025. IFAAMAS. doi:10.5555/3709347.3743564.

5 Amit K. Chopra and Munindar P. Singh. Cupid: Commitments in relational algebra. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI)*, pages 2052–2059, Austin, Texas, January 2015. AAAI Press. doi:10.1609/aaai.v29i1.9443.

6 Amit K. Chopra and Munindar P. Singh. Custard: Computing norm states over information stores. In *Proceedings of the 15th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1096–1105, Singapore, May 2016. IFAAMAS. doi:10.5555/2936924.2937085.

7 Amit K. Chopra, Leon van der Torre, Harko Verhagen, and Serena Villata. Normative multi-agent systems (Dagstuhl seminar 15131). *Dagstuhl Reports*, 5(3):162–176, 2015. doi:10.4230/dagrep.5.3.162.

8 Civil Resolution Tribunal of British Columbia. Moffatt v. Air Canada, 2024 BCCRT 149 (CanLII), February 2024. Accessed 2025-09-02. URL: https://canlii.ca/t/k2spq.

9 Dov Gabbay, John Horty, Xavier Parent, Ron Van der Meyden, Leendert van der Torre, et al. *Handbook of Deontic Logic and Normative Systems, Volume 1*. College Publications, 2013.

10 Dov Gabbay, John Horty, Xavier Parent, Ron Van der Meyden, Leendert van der Torre, et al. *Handbook of Deontic Logic and Normative Systems, Volume 2*. College Publications, 2021.

11 LangGraph. LangGraph: Building language agents as graphs, December 2024. Accessed 2024-12-05. URL: https://langchain-ai.github.io/langgraph/.

12 Adrian Lee, Kip Martin, and David Yockelson. Stop agent-washing: Differentiate with human-centric agentic experiences. Technical report, Gartner Research, 2025. URL: https://www.gartner.com/en/documents/6819934.

13 Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents, 2023. URL: https://arxiv.org/abs/2308.03688, arXiv:2308.03688.

14 Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general ai assistants, 2023. URL: https://arxiv.org/abs/2311.12983, arXiv:2311.12983.

15 Munindar P. Singh. Information-driven interaction-oriented programming: BSPL, the Blindingly Simple Protocol Language. In *Proceedings of the 10th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 491–498, Taipei, May 2011. IFAAMAS. doi:10.5555/2031678.2031687.

16 Munindar P. Singh. Semantics and verification of information-based protocols. In *Proceedings of the 11th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1149–1156, Valencia, Spain, June 2012. IFAAMAS. doi:10.5555/2343776.2343861.

17 Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21:1–21:23, December 2013. doi:10.1145/2542182.2542203.

18 Munindar P. Singh. Consent as a foundation for responsible autonomy. *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 36(11):12301–12306, February 2022. Blue Sky Track. doi:10.1609/aaai.v36i11.21494.

19 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models, 2023. URL: https://arxiv.org/abs/2210.03629, arXiv:2210.03629.

5.3 Conceptual Issues Regarding Policy Modelling and Reasoning in Sociotechnical Systems

Cristina Baroglio (University of Turin, IT), Mehdi Dastani (Utrecht University, NL), Frank Dignum (University of Umeå, SE), Beishui Liao (Zhejiang University, CN), Vivek Nallur (University College Dublin, IE), Judith Simon (Universität Hamburg, DE), Sz-Ting Tzeng (University of Umeå, SE), Harko Verhagen (Stockholm University, SE), and Jessica Woodgate (University of Bristol, GB)

5.3.1 Introduction

This working group focused on conceptual issues perceived to relate to policy modelling and reasoning in STS. The terminology problem was first discussed, as different stakeholders view STS differently which causes a lack of consensus in the use and misuse of a system. Difficulties in ascertaining what constitutes misuse of an STS, or when misuse is occurring, present barriers to deciding and enforcing effective regulation. The discussion went on to apply conceptual issues to an example of an educational STS, which encompasses dynamic change of goals and behaviour, influencing the technological decisions and process that must be made.

5.3.2 Boundaries, Mental Models, Agency

Efforts towards regulation can only address reasonably foreseeable misuse whilst leaving room for innovation. Yet, human behaviour is extremely hard to predict. In an STS composed of social (humans and organisations) and technical (data and devices) tiers, there will necessarily be a limit to the extent of what the human tier can understand, and what the technical tier can represent.

The lack of fixed boundaries in STS terminology leads to a lack of consensus about the use and misuse of a system. Technology-only solutions to STS misuse typically result in a blunt instrument approach, with the social components adopting, rejecting, complying, or violating the assumptions and norms (standards of expected behaviour [5]) of the STS. Humans create a mental model of the system (both technical and inter-related processes), yet it is unclear if the technical tier does, or is able to, represent such a model. Resulting problems from this include:

Deception: Both intended and unintended

Trust: Both deserved and undeserved

Goal and value alignment: Humans tend to assume that the technical system *ought* to have the humans' best interest at heart.

Can these problems be fixed with a notion of *group agency*? Agency ought to lead to responsibility, and hence group agency makes explicit the notion of shared responsibility [3]. Understanding how group agency emerges necessitates defining group membership, structure, and behaviour.

5.3.3 Normativity

Within an STS, stakeholders attempt to achieve normative goals, wherein normativity refers to something desirable. Normative goals may be imbued through law, social norms, or moral norms, and explicit or implicit incentives may be applied to influence behaviour and encourage stakeholders to comply with particular norms. The nature of an STS is an ongoing and dynamic process, where humans influence systems and those systems in turn influence human behaviour. Therefore, when a choice is made to encourage a normative goal in a system (e.g. installing smoke detectors to stop people smoking indoors), we can expect that people will adapt in response to that choice (e.g. by not smoking or by taking the batteries out of the smoke detector). Norms thus embody a dual role, effecting expected changes and catalysing unexpected responses to those changes. The technological choices we make in pursuit of normative goals will never be able to capture the whole space of normativity as people will loopholes and behave in unforeseen ways. Smoke alarms are installed to promote safety, but some people deviate by covering the alarm up or removing the power.

Communication fills the gap between the technology itself, and the normative goal aimed at, through sanctions (positive or negative reactions to approved or disapproved behaviour [2]) or other signals [4]. For example, speed signs may be put up to make people slow down outside a school. However, interventions may change over time as the system responds. If people do not follow the signs, the intervention may need to be fortified such as by putting in a speed bump.

Technological interventions to promote normative goals raise questions about intrinsic autonomy and deception. Whether a deception is (un)acceptable may depend on the intentions and reasons behind it, as well as the awareness of the subject that they are being deceived. On the one hand, end users should be aware of the limits of the technology they are engaging with. For example, some large LLMs now have a disclaimer that output may be misleading or false which alerts the user to the fact that LLMs are fallible. On the other hand, if the communication about the limits of a technology is incomplete, misunderstandings may arise. As the role of LLMs in everyday life increases (e.g. as the first result of a search engine), increasing exposure to the disclaimer that LLMs give inaccurate responses may reinforce the belief that people cannot trust the information they encounter, adversely affecting the way they interpret other sources of information such as the news. LLM outputs are deceptive in part because there is a detachment from the labelling of training data and the uncertainty of the output. Where a human would convey uncertainty in their communication, LLMs do not.

Whilst there are important limits to technological interventions to achieve normative goals, some people will still have a tendency to overestimate the capabilities of and defer to technology. It is thus key that efforts are made towards designing technology in ways that both benefit the system and acknowledge its limits.

5.3.4 Trust

The typical (philosophical) conception of trust is understood as:

A *trusts* B *with regard to* X

Trustworthiness has moral and an epistemic requirements:
Epistemic: To know whether something is competent and know the limits of that competence
Ethical: To know if the entity being trusted is honest and benevolent

It is critical to differentiate between `trust` and `trustworthiness`. Trust is an intentional stance taken by an entity, while trustworthiness is a property or a relation which perhaps can be measured. Trust in technology often reduces to reliance. Trust need not be compositional, *i.e.*, **A** can trust **B** without trusting all the constituent parts of **B**. In the application of system design, trust functions as a social enabler, allowing social entities to function without requiring third-party guarantees.

5.3.5 Value Alignment

There is some debate over whether global values really exist. Within this debate, the following are some important questions:
Understand: **How** do we define the value?
Define: Who decides **what** the relevant values are?
Alignment: A **joint-expectation** on what actions are plausible, and which ones are good
Reconcile: How do we decide whether the components of an STS are value-aligned, when there are **differences** between parts?

One possible characteristic of value is that it forms the criteria for comparing situations. A norm can be understood as a preference relation that pushes an entity towards actions leading to a situation where a value is upheld. This implies that the norms adopted influence the kinds of values that can be upheld. A value-conflict can therefore be defined as differences in estimates of goal states/situations, based on actions pushed by the norms. *Can value-alignment be defined as the absence of value-conflict?*

An alternate definition of value-alignment is that if two entities **A** and **B** take the same action, given the same context, one can form a belief that A and B are value-aligned. The presence of value-alignment creates the presence of an in-group *vis-a-vis* the out-group that (by definition) does not align with the same values. Members of the in-group have a common ordering over shared values.

It is not very well-understood how one should choose between values. Values that an entity is unwilling to negotiate on, are called non-negotiable values. These could also be viewed as sacred values.

5.3.6 Observation and Verification of Values

Designing a system given a set of shared values, considerations include how to encode the requirements that flow from those values, and discerning the evidence or data that is needed for the design of systems. It is unclear whether a value is something that can be observable

by perception, or is something that is completely cognitive. Agent architectures that could represent values include:

- Reactive
- Deliberative (*e.g.*, Belief Desire Intention, BDI)
- Symbolic
- Sub-symbolic

For sub-symbolic architectures, if the mechanism is very good then representation becomes invisible. Sub-symbolic reflects values present in the training data; values can be represented using reinforcement learning from human feedback (RLHF) [1]. Possible mechanisms for collating the training data for sub-symbolic approaches include social networks and purposeful collection.

Exploring whether values can be mixed in an STS, and if it is possible to validate a consistent mixing of values, is a gap in research. Possibly, validation of value-mixing can only be achieved in specific forms of value-failure (e.g., discrimination based on race or gender).

5.3.7 Argumentation-Based Methodology for Representing and Reasoning about Policies, Norms, Values, Disobedience, and Causality

Policies, norms, and values exhibit inherent potential conflicts and dynamic evolution within STS. Formal argumentation establishes a principled methodology for representing and reasoning about these entities, extending to norm violations and causal relationships. The integrated framework comprises:

Unified Representation: Norms, policies, and their violation conditions (disobedience) are modelled uniformly as defeasible rules. Value sets annotate norms, while priority relations operate over rules or their value associations

Conflict & Violation Resolution: Argumentation resolves conflicts between competing norms/policies and adjudicates norm violations. Embedded in belief-desire-norm-policy-intention (BDNPI) architectures, this enables autonomous agents that reconcile policy guidance with normative constraints, including disobedience detection and sanction reasoning

Dynamic Adaptation: Runtime modification of norms, policies, and violation thresholds is facilitated through argumentation revision mechanisms, maintaining system consistency during change

Causal Attribution: Argumentation frameworks can be extended to integrate causal inference for modelling responsibility attribution. This enables: (1) trust establishment via transparent causality chains; (2) precise accountability assignment for norm violations; and (3) root-cause analysis of system failures across multi-stakeholder interactions

Computational Viability: Locality-driven computation and modular design overcome complexity barriers, ensuring efficient handling of dynamic rule updates and causal reasoning

Neuro-Symbolic Integration: LLMs transform natural language norms/violation clauses into formal defeasible rules, while argumentation provides rigorous conflict and causality resolution – enabling hybrid reasoning unattainable by monolithic approaches

5.3.8 Governance of an STS

In governing an STS, it is important to acknowledge that any STS that needs to adapt to changing participants and technologies will become a complex system. If such a system is to be adaptable to change, then it needs to be governed rather loosely. A general guiding principle could be to *identify the forces that move the STS to some attractor point, and try*

to govern those forces. Diving deeper, we discussed the reasoning model required of agents, in an agent-based simulation of an STS. Some relevant questions include:

- How deep should agent reasoning attempt to go? Would reactive agents with appropriate calibration suffice? Or are BDI agents necessary?
- Can the two agent architectures be systematically bridged?
- What are the HPC requirements that agent-based models should consider, while being designed?
- Are there integrative techniques for merging sampling data, with deeper survey-type models?
- Which communities need to attempt integrating their techniques? (e.g., computer science, social psychology, behavioural economics)

It was recognised that we need to have some reference problems the community could attempt to solve. One suggestion was that the community agrees on a small set of small-but-complex systems, and attempt to answer each of the above questions systematically. This could possibly involve holding an iterated competition that starts from the same codebase each time it is held. Learnings from multi-year competitions could be applied to bigger and more realistic STS.

5.3.9 Policy Modelling

Policy modelling is a multidisciplinary field that synthesises insights from behavioural sciences, social simulation, reasoning, humanities, and history. In policy modelling, data doesn't mean reasoning. Agent behaviour for simulation can come from data, but it can also come from asking what people do and what is important to them. The latter covers humans' reasoning process.

In simulation, the input for how (human) agents reason can rely on basic assumptions (e.g. norm-obeying willingness based on political affiliation), theories (e.g. social-psychology theories), and actual data collected (e.g. using various social science methodologies). It is especially interesting and relevant for policy modelling in STS how (if at all) humans perform normative reasoning when making decisions. While there seems to be an agreement that integrating the different approaches would be beneficial for the fruitfulness of simulation, it is far from obvious how the integration should happen. Integrating diverse approaches is hence an open challenge for the community.

5.3.10 Education as a Use Case

Building on abstract conceptual ideas, the discussion transitioned to consider an educational STS. Education presents a useful case study of an STS that is dynamic and complex. In particular, we considered teaching computer science at university. Education and learning is a domain where behavioural change, due to technological decisions and processes, might be both short and long-term. An educational STS is a (as, we imagine, many STS are) complex system. The complexity means that we have no systematic way of determining which point of intervention is the best, or if the intervention will cause the future to be as we envisioned. Important questions discussed included:

- What are the challenges of an STS where each student might rely on one or more GenAI systems (supplied or not supplied by the university)?
- Would GenAI systems be considered as personal agents? How should GenAI be integrated into an STS?
- What are, or ought to be, the learning goals?

- How do we reach learning goals when tools change substantially, explicitly accounting for skills that we believe could be lost?
- How do we evolve or change goals over time? For example, reading and writing are currently considered important. Will they continue to be so? Are they important for critical thinking?
- Are we using tools as assistive technologies, or skill replacement technologies?
- What happens when our values and goals change? Some jobs/careers may diminish in importance when tasks are automated, so that they are not coveted anymore.

5.3.10.1 What Topics to Teach

The approach to programming changed rapidly in the presence of GenAI systems, does it make any sense to keep on teaching programming as in the past. So, do we need courses on programming and/or teach on the use of AI programming "tools" and have students (inter)act with the tools – pair-programming in a human-AI team as the future of software development. Some drawbacks are:

What to teach: Programming with Gen AI is a loop where the human uses the system to revise a drafted program improving it until he/she is satisfied with it. But, the roles are not equal: the human has responsibility over the product as well as the ownership in terms of copyright, for this reason he/she should have the ability to evaluate the program that is being built

Normative compliance: The GenAI system often produces over-complex code, difficult to understand or, more in general, not respecting some general norms, policies, or good practices we would like to be respected

Personal development: Companies want to hire persons that have high level skills, and who can work with abstractions so that they can understand if and how things fit together. Attractive candidates do not try to solve a problem on their own but rather know when and how to ask things to teammates. So, how does programming fit in here? Perhaps we should put forward more general learning outcomes that encompass critical thinking and analytical skills, as well as understanding loops and recursion

5.3.10.2 How to Teach

Suitable approaches could take inspiration from "the Amazon method", which involves starting with a mandatory brainstorming session and is followed by an idea collecting session. Teaching could use design thinking as a model, or focus on more student-centred methods and flipped classroom-inspired teaching. Currently, most programming courses have lab sessions with teaching assistants where students get help "on the fly" in constructing their program. Teaching assistants are increasingly receiving student requests to explain code that works when executed, but students do not understand why because it was produced with the help of an LLM. This is connected to the previous item – the ownership and understanding of what the code does are not matching. Thus, different strategies for teaching and for assessing what has been learned should be developed.

5.3.10.3 How to Assess

As LLMs have destroyed the essay (and in some discussions, the bachelor or masters thesis has also been declared dead), the assessment of programming skills may need to adapt. In an LLM-supported programming curriculum, it may not be all that important to check that a student wrote his/her code without help, as in companies they will probably use LLMs

and work in teams, so it is important that they practice with such tools. Yet, students should be able to explain the code, the choices made, why they are good, their limits, and so on. Similarly, it is important that students can review code written by someone else. More appropriate assessment could thus take the form of oral on-the-spot explanations of code. Such assessments should take place in a controlled environment.

We observe that when writing text, LLMs often mark a particular words or ways of phrasing things as a mistake, forcing a "standardised" way of writing. Similar instances may happen with writing code, limiting the exploration of learners. This is the issue of sausage production: GenAI tools repress individual expressions and start from "the mean", which results in everything becoming gray (and correcting correct items to incorrect or bland ones as a consequence).

On the whole, we would like students to learn to be in control of Gen AI systems when using them to produce code, they should be able to review code but for this aim they should know and have practice on how to code. They should be aware of the strengths and limits of such systems, and not be over-reliant on them.

5.3.11 Conclusions

Loose boundaries of terminology related to STS confuses how to define (mis)use of a system, which in turn makes STS difficult to regulate. Normativity, trust, and values are fundamental to STS, yet there are important challenges with attempts to granulate these concepts into entities that can be encoded. Formal argumentation could be useful to represent and reason about the conflicts associated with policies, values, and norms. An educational STS provides a helpful use case to explore how the conceptual issues discussed actualise and how challenges could be addressed.

References

1 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 1–9. Curran Associates, Inc., 2017.

2 Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review (KER)*, 31:142–166, March 2016.

3 Raimo Tuomela. Joint Intention, We-Mode and I-Mode. *Midwest Studies in Philosophy*, 30(1):35–58, 2006.

4 Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. Norm enforcement with a soft touch: Faster emergence, happier agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1837–1846, Auckland, May 2024. IFAAMAS.

5 Georg Henrik Von Wright. Deontic logic: A personal view. *Ratio Juris*, 12(1):26–38, 1999.

5.4 Harmonising constraint and policy languages for the use by autonomous agents

Nicoletta Fornara (USI – Lugano, CH), Beatriz Esteves (Ghent University, BE), Sebastian Neumaier (FH – St. Pölten, AT), Victor Rodriguez Doncel (Polytechnic University of Madrid, ES), Simon Steyskal (Siemens AG – Wien, AT), Rigo Wenning (W3C / ERCIM, FR), and Antoine Zimmermann (Ecole des Mines – St. Etienne, FR)

5.4.1 Introduction

The subgroup of the seminar addressed questions arising from the use of constraint and policy languages. Presentations of the ODRL Policy language and the SHACL constraint language were given. The group discussed challenges around using SHACL and ODRL, how they relate, how they differ, and how they can be used in combination. We found that SHACL can at least serve to disambiguate the expression of constraints in ODRL. Both languages are W3C Recommendations.

A standardisation strategy discussion was triggered. It included the suggestions to amend the respective standards to ease the combination of ODRL and SHACL. Standardisation gaps were identified and the group discussed the formal semantics of ODRL and how to update the current W3C Recommendation. Given the work on RDF 1.2, the group explored how to use the new possibilities for annotation of data coming with those updates.

Web agents and the Web Agents Community Group and their work were presented. We discussed how policy and constraint languages can help in an agent scenario. New challenges stemming from web agents for policy and constraint management were identified. Annotation and protocol issues were at the core of our discussion.

A *policy language* is a formal or semi-formal language used to express various types of rules (i.e., permissions, obligations, and prohibitions) that govern access and usage of resources or regulate behaviour or state of affairs in open distributed systems. It typically includes constructs to define subjects, actions, objects, and conditions under which actions or states are allowed or denied. Policy languages are used in a wide range of domains, including access and usage control, data usage, privacy, and digital rights management.

5.4.2 Background

5.4.2.1 From Rights Expression Languages to Policy Languages

ODRL 1.1 (Open Digital Rights Language) was introduced as a *Rights Expression Language* (REL) designed to represent statements about the usage rights of digital content. Published as a W3C Note in 2002 [14], it emerged in the context of the growing interest in Digital Rights Management (DRM) systems during the early 2000s. These systems required standardised means to express permissions and conditions associated with digital assets.

ODRL 1.1 was based on XML and offered a structured and extensible way to define rights and conditions. Its clean and modular specification contributed to its adoption in several industry sectors. Notably, ODRL 1.1 was incorporated into the Open Mobile Alliance (OMA) DRM specifications [23], where it served as the core rights language for managing usage rights on mobile devices. This adoption demonstrated the practical applicability and interoperability of ODRL 1.1 across platforms and devices even without having a formal semantics, the specification was clear for everyone and the interoperability problems were minimal or non-existent.

The XML-based ODRL 1.1 was later replaced by a more expressive RDF-based version, formalised in the ODRL 2.2 ontology [22]. This revision was not merely syntactic; it introduced significant enhancements in the language's expressive power. In particular, the core concept of a *rights expression* was replaced by the more general notion of a *policy expression*. While `odrl:permission` had been sufficient for access control scenarios, ODRL 2.2 introduced additional constructs such as `odrl:prohibition` and `odrl:obligation` (also referred to as duties), allowing for richer and more nuanced representations of policy constraints. These expanded capabilities significantly increased the applicability of the language across broader domains. However, they also introduced greater interpretive complexity and potential ambiguity in how policy expressions should be understood and enforced.

5.4.2.2 The ODRL 2.2 Policy language

The Open Digital Rights Language (ODRL) [15] is the policy language for the Web specified by the W3C[4] for expressing permissions, prohibitions, duties, and restrictions associated with digital content. ODRL is used in contexts such as rights management, licensing, and, more recently, data sharing agreements [26]. It is specified in two W3C Recommendations: the Information Model [15] and the Vocabulary & Expression [16], which are primarily text documents, but also include an OWL ODRL Ontology[5] and a number of profiles and supplementary documents, such as [17, 27, 6, 13, 21].

The syntax of valid ODRL policies must therefore conform not only to the rules formally specified in the OWL ontology, but also to additional constraints described in natural language in the specification. For example, the OWL ontology states that the domain of the `odrl:constraint` property must be either an `odrl:Policy` or an `odrl:Rule`. However, the ontology does not impose that a policy must contain at least one rule, although this requirement is expressed in the textual specification [15, Section 2.1]. Many such textual constraints can be represented as SHACL shapes used for validation,[6] but these rules are not formally standardised. Thus, it is fair to say that ODRL is a *semi-formal language*.

5.4.2.3 Formal Semantics for ODRL

The ODRL specification refers to an "ODRL Validator" –a system that checks the conformance of ODRL Policy expressions – but its behaviour is even less precisely defined. Providing a fully formal semantics for this software component would bring significant interoperability benefits. This need led to the creation, in 2021, of a draft *Formal Semantics for ODRL* report [10], which is still under development and was a topic discussed at the seminar. The semantics of ODRL 2.2 can be specified in a *declarative* format (as in [1]) or can be specified in an *operational* format by providing an algorithm for translating ODRL policies into another language that has a formal semantics, for example SPARQL.

5.4.2.4 Constraint languages in the Web

A *constraint language* is a language that can be used to express conditions or integrity constraints over data structures. These constraints can be used for validation, consistency checking, or to define requirements that data must satisfy. Constraint languages are often declarative and are used to ensure that (structured) data adheres to specified rules, independently of any procedural behaviour.

---

[4] World Wide Web Consortium, `https://www.w3.org/`
[5] `https://www.w3.org/ns/odrl/2/`
[6] ODRL Implementation, `https://odrlapi.appspot.com/`

SHACL (Shapes Constraint Language) [20] is the W3C recommendation used to express constraints over RDF data. It allows for the specification of conditions that RDF graphs must satisfy and is commonly used for data validation. SHACL 1.2, a Working Draft as of July 2025 [19], extends the original specification with several significant features aimed at increasing expressivity, usability, and integration with RDF and SPARQL.

In the context of ODRL, SHACL has been used to encode additional constraints that are specified in the natural language part of the ODRL specification but not captured by the OWL ontology.

5.4.2.5 Automated Regulatory Compliance

Automated regulatory compliance faces several fundamental challenges. These potentially stem from the inherent complexity of legal texts and the diversity of technical systems that the legal text aims to govern. Regulations are typically written in natural language, using ambiguous terms, implicit assumptions, and context-depended information. Translating the natural language legal provisions into a formal language like SHACL or ODRL is not an unambiguous operation. Translating a legal text may not always result in the same formal language file. The result of a transformation of a legal compliance requirement is thus necessarily and always an interpretation of the law creating that compliance requirement.

On the other hand, compliance is not just about the implementation of static checks (e.g., the completeness of meta-information); many obligations depend on dynamic and evolving system behaviour (for instance, consider updates in the provenance of training data, or in the deployment process of AI systems). The data side of things may not be uniform either. So the compliance checking needs to be done over a very heterogeneous data landscape. In fact, the constraint file serves as a first filter that searches for contextual graph artifacts matching the constraint shape that resulted from the interpretation of law.

Currently, there is a need for better standardised mappings between legal norms and technical artifacts that potentially lead to ad-hoc, domain specific implementations. In the future, a law establishing a new compliance requirement may choose to create that SHACL constraint file and the subsequent action to accomplish as an annex to the actual law. In this case, the legislator does the translation himself. Consequently, the constraint file will participate in the normative and authoritative value of the legislator. As a consequence, matching the constraint will then mean the official recognition of compliance. A system of automatic compliance testing and confirmation would appear.

Looking at the concrete case of using ODRL to represent policies for aiding with regulatory compliance, it has been concluded that ODRL is not fit for purpose to represent legal concepts, such as purposes or legal grounds under which data can be accessed or used, as it does not contain such concepts in its vocabulary [9]. As such, one can make use of its profile mechanism to extend the ODRL vocabulary with concepts for representing contextual information relevant for legal compliance. Relevant work [5] in this area has been explored in the context of the SPECIAL project, which used ODRL constructs and legal concepts from vocabularies established in the context of this project to support regulatory compliance checking of business policies.[7] The SPECIAL project also launched the work of the W3C's Data Privacy Vocabularies and Controls Community Group (DPVCG).[8] The mission of this group is to develop specifications for representing machine-readable metadata about the use

[7] SPECIAL project homepage, `https://specialprivacy.ercim.eu`, retrieved 3 July 2025.
[8] Data Privacy Vocabularies and Controls Community Group homepage, `https://www.w3.org/community/dpvcg/`, retrieved 3 July 2025.

and processing of personal and non-personal data, as well as about technologies that use such data, in a jurisdiction-agnostic manner, and also create extensions to these specifications for concrete regulations, such as the European General Data Protection Regulation (GDPR) or the AI Act. The core specification, the Data Privacy Vocabulary (DPV) [25, 7] includes taxonomies to represent information about entities and legal roles, purposes and processing operation concepts, data and personal data, rights, risks, contextual information about processing operations, such as storage conditions or the scale of processing, technical and organisation measures, legal bases, and location and jurisdiction terms. Hence, by using both DPV and ODRL, i.e., making use of ODRL's profile mechanism, one can model policies using ODRL's model while using DPV's terms to refer to legal concepts [8, 24]. Furthermore, the previously cited publications are currently being used, in a joint effort by the ODRL and DPV communities, as the basis to create an official DPV-ODRL profile[9], and a guide document for using this profile[10]. As such, DPV is a promising approach to tackle regulatory compliance, when used with a policy language such as ODRL. It is still continuously being maintained and updated with new requirements coming from newly-enforceable laws, such as the European Health Data Spaces Regulation or the AI Act.

5.4.2.6 Governing agents on the Web using policies

In the context of autonomous agents on the Web, expressing policies formally is crucial, yet still a key challenge [18]. Web-based systems may be open to new, previously unidentified agents that must be guided by way of systematic formal knowledge upon which they can carry out logical deductions. On the Web, agents can autonomously discover more information about the resources they deal with thanks to the hypermedia dimension of REST-based infrastructure. Following a link, a Web agent can navigate from an object description to norms associated with it, to policies, and more [4].

If agents are implemented such that they can automatically follow policy descriptions, then they can optimise the utilisation of the Web resources, including–potentially–getting the assistance of other agents that have observable presence on the same Web platforms [28]. On the contrary, if the agents do not obey the rules imposed by policies, they may be driven away by either the Web platform that hosts the resources, or by other autonomous Web agents that have norm-enforcing role.

In such scenarios where artificial agents must adhere to policies as strictly as possible, and connect policies to resource descriptions as well as possibility of interactions, it is convenient to rely on knowledge graphs as the underlying data model and technology. As argued in [2], knowledge graphs are key enablers of autonomy in relation to all aspects of autonomy, although challenges relating to governing agents remain open.

Because of the open challenges in designing Web-based MAS, academic researchers and enterprise practitioners are crossing their views and insights over a W3C community group exploring Web Agent technologies, including recent advances in LLM-based Agentic AI.[11] The group identified a strong interest by academia and corporation alike to devise standardised interaction protocols, where norms and policies represent a key dimension for supporting the governance of agents on the Web [3, Section 10].

[9] Mapping from DPV to ODRL (Draft Community Group Report), `https://w3id.org/dpv/mappings/odrl`, retrieved 4 July 2025.
[10] Guide for using DPV with ODRL (Draft Community Group Report), `https://w3id.org/dpv/guides/dpv-odrl`, retrieved 4 July 2025.
[11] W3C Autonomous agents on the Web community group, `https://www.w3.org/community/webagents/`

5.4.3 Main Discussions

The ODRL Evaluator is designed to produce conclusions by performing logical reasoning over ODRL policies, an evaluation request, and a state of the world. Various approaches have been explored to achieve this, including mappings to finite state machine systems, formal logic systems, Answer Set Programming (ASP), and logic programming languages such as Prolog.

In this seminar, the potential use of SHACL has been explored. While SHACL has demonstrated effectiveness in validating the syntactic correctness of ODRL policies, it also shows promise as a possible reasoning engine underlying the ODRL Evaluator.

5.4.3.1 Expressing ODRL Constraints with SHACL

ODRL constraints, such as a rule limiting usage to ten hours, are typically expressed using the ODRL vocabulary, for example by defining a constraint with odrl:leftOperand, odrl:operator, and odrl:rightOperand. The same restriction can also be validated operationally with SHACL: a sh:PropertyShape can be defined on the property that records actual usage time, here ex:totalUsageTime, with sh:maxInclusive "PT10H"8sd:duration. This SHACL shape ensures that any recorded usage value exceeding ten hours will be flagged as invalid, making the policy both human-readable in ODRL and machine-checkable through SHACL.

```
[ a odrl:Constraint ;
  odrl:leftOperand odrl:meteredTime ;
  odrl:operator odrl:lteq ;
  odrl:rightOperand "PT10H"^^xsd:duration
]
```

Figure 6 ODRL Constraint.

```
[ a odrl:Constraint ;
  odrl:leftOperand odrl:meteredTime ;
  odrl:operator odrl:lteq ;
  odrl:rightOperand "PT10H"^^xsd:duration
]
```

Figure 7 SHACL Shape.

Figure 8 ODRL Constraint → SHACL Shape.

An important direction for future work is the definition of a general mechanism for mapping ODRL constraints to SHACL constraints, taking into account the variety of left operands defined in the ODRL Vocabulary. A key difficulty lies in bridging the abstract semantics of ODRL's left operands (e.g., `odrl:meteredTime`) with the concrete properties used in specific datasets or implementations (e.g., `ex:totalUsageTime`).

5.4.3.2 The challenge of operationalising compliance of AI systems

Automated regulatory compliance, particularly in the context of the European Union's AI Act, faces challenges in mapping high-level legal requirements to concrete technical artifacts [11]. The key difficulty often lies in the insufficient or fragmented information available across the AI lifecycle, which complicates interpretation and verification of transparency and traceability of AI systems.

Compliance frameworks for AI, such as the framework envisioned in the projects CERTAIN[12], HARNESS[13] or GLACIATION[14] aim to address this by formalising metadata and

[12] CERTAIN project homepage, https://certain-project.eu/, retrieved 4 July 2025.
[13] HARNESS project homepage, https://harness-network.eu/, retrieved 4 July 2025.
[14] GLACIATION project homepage, https://glaciation-project.eu, retrieved 4 July 2025.

aligning it with regulatory requirements [12]. However, integration of these representations require clear machine-processable formalisation of policies, which is a non-trivial task due to semantic ambiguity and evolving standards.

A promising approach to mitigate these issues is the current developments of SHACL: the use of SHACL as a validation mechanism to enforce structured constraints on policy-relevant metadata, enabling systematic compliance checks against legal requirements. For instance, the EU AI Act's Annex IV requires providers of high-risk AI systems to maintain technical documentation, such as performance metrics of an AI system. SHACL can be used to validate completeness and structural requirements – given that the underlying information is sufficiently formalised.

5.4.4 Proposed approaches

5.4.4.1 RDF 1.2 Triple Terms

The integration of RDF 1.2 capabilities into ODRL policy evaluation opens new possibilities for creating more robust and transparent policy enforcement mechanisms. Beyond the provenance and duty fulfilment scenarios outlined above, RDF 1.2 *annotations* can address several critical challenges in policy evaluation and compliance monitoring.

Policy State Tracking

A fundamental challenge in ODRL policy evaluation is maintaining state information across multiple policy interactions. Traditional approaches often rely on external state management systems, creating potential inconsistencies between the policy representation and its execution state. RDF 1.2 triple terms and reifiers together with Turtle 1.2's suggested Annotation Syntax [15] enable embedding state information directly within the policy graph itself.

This approach ensures that policy state remains synchronised with the policy definition, enabling more reliable policy evaluation and reducing the complexity of external state management systems.

Explainable Policy Decisions

Policy evaluation often involves complex reasoning processes that can be difficult to trace and explain. RDF 1.2 annotations provide a mechanism for embedding explanation trails directly into policy evaluation results, supporting both accountability and debugging requirements.

Such annotated evaluation results provide transparency into the decision-making process, enabling stakeholders to understand why specific policy decisions were made and facilitating trust in automated policy enforcement systems.

Temporal Policy Evolution

Policies often evolve over time, and maintaining a clear audit trail of policy changes is crucial for compliance and governance. RDF 1.2 annotations enable the embedding of version control information directly within policy definitions, creating self-contained policy evolution histories.

[15] Turtle 1.2 Annotation Syntax `https://www.w3.org/TR/rdf12-turtle/#annotation-syntax`

```
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <http://example.com/policy/> .

:policy003 a odrl:Agreement ;
 odrl:permission [
  odrl:target :document456 ;
  odrl:action odrl:read ;
  odrl:constraint [
   odrl:leftOperand odrl:count ;
   odrl:operator odrl:lteq ;
   odrl:rightOperand "5"^^xsd:integer
   # --- State Annotation ---
   {| :currentCount "2"^^xsd:integer ;
     :lastAccessed "2025-07-02T14:30:00Z"^^xsd:dateTime ;
     :accessedBy :user456
   |}
  ]
 ] .
```

Figure 9 Policy with embedded state tracking.

```
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <http://example.com/policy/> .

# Policy evaluation result
:evaluation001 a :PolicyEvaluation ;
 :evaluates :policy003 ;
 :result :denied
 {| :reason "Usage count limit exceeded" ;
   :evaluationTime "2025-07-03T10:15:00Z"^^xsd:dateTime ;
   :evaluatedConstraints (:constraint001 :constraint002) ;
   :failedConstraint :constraint001
 |} .
```

Figure 10 Policy evaluation with explanation annotations.

This approach enables policy systems to maintain comprehensive change histories without requiring external versioning systems, supporting both compliance requirements and operational transparency.

5.4.4.2 Validating Triple Terms with SHACL 1.2

A reified statement is an RDF triple (subject-predicate-object) turned into a resource that can itself be the subject of other statements. SHACL 1.2 supports reified statements by providing enhanced validation capabilities for metadata attached to statements, including support for RDF-star syntax and traditional RDF reification patterns. This allows SHACL shapes to validate not only the original triple but also the properties describing when, how, or by whom the statement was made.

```
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <http://example.com/policy/> .

:policy004 a odrl:Agreement ;
 odrl:permission [
  odrl:target :sensitiveData ;
  odrl:action odrl:use ;
  odrl:constraint [
   odrl:leftOperand odrl:purpose ;
   odrl:operator odrl:isA ;
   odrl:rightOperand :researchPurpose
  ]
 ]
 {| :version "2.1" ;
   :previousVersion :policy004_v2.0 ;
   :modifiedBy :admin_bob ;
   :modificationDate "2025-06-15T16:45:00Z"^^xsd:dateTime ;
   :changeReason "Updated purpose constraints per new research guidelines"
 |} .
```

Figure 11 Policy with version history annotations.

For example, a SHACL shape can require that every occurrence of an odrl:permission triple is accompanied by provenance metadata describing its version and possible link to a previous version. In the example below, the shape ex:ProvenanceShape specifies that reified permission statements must include at most one :version (typed as an xsd:date) and at most one :previousVersion (an IRI pointing to an earlier policy). The property shape ex:PermissionShape then declares that any odrl:permission triple must be reified and must conform to that provenance shape. In this way, SHACL 1.2 ensures that policies are not only structurally correct but also carry the necessary temporal and versioning information for accountability.

5.4.5 Conclusion

The group intends to dig deeper into the topic of ODRL semantics and using SHACL constraints. The results are such that a scientific article seems at reach. On the basis of that article, future work on ODRL will be suggested. This may lead to the creation of a new ODRL Working Group in W3C that would create a simplified profile of ODRL 2.2, and also do some maintenance work (i.e., correcting some issues in ODRL 1.1's W3C Note) and allow for the use of SHACL in the constraint element of ODRL. It could also further explore how to create Policy annotations of data using RDF 1.2. To kick off this initiative, a Workshop needs to be organised.

References

1 Piero Bonatti, Nicoletta Fornara, and Andreas Harth. Towards a Formal Semantics of the Open Digital Rights Language (ODRL 2.2). In Marta Sabou, Andreas Harth, Pasquale Lisena, Edward Curry, Bohui Zhang, Reham Alharbi, Yuan He, Georg Rehm, Sonja Schimmler, Stefan Dietze, Natalia Manola, Andrea Cimmino, Nicoletta Fornara, Víctor Rodríguez-Doncel, John Domingue, Achim Rettinger, Damian Trilling, Marko Grobelnik, Claudia d'Amato, Valeria Fionda, Ilaria Tiddi, and Gabriele Tolomei, editors, *ESWC 2025*

```
ex:ProvenanceShape
  a sh:NodeShape ;
  sh:property [
    sh:path :version ;
    sh:datatype xsd:date ;
    sh:maxCount 1 ;
  ] ;
  sh:property [
    sh:path :previousVersion ;
    sh:nodeKind sh:IRI ;
    sh:maxCount 1 ;
  ] .

ex:PermissionShape
  a sh:PropertyShape ;
  sh:path odrl:permission ;
  sh:reifierShape ex:ProvenanceShape ;
  sh:reificationRequired true .
```

Figure 12 SHACL 1.2 proposes constraint components for validating reifiers.

Workshops and Tutorials Joint Proceedings, volume 3977 of *CEUR Workshop Proceedings*. Sun SITE Central Europe (CEUR), June 2025. URL: http://ceur-ws.org/Vol-3977.

2 Jean-Paul Calbimonte, Andrei Ciortea, Timotheus Kampik, Simon Mayer, Terry R. Payne, Valentina Tamma, and Antoine Zimmermann. Autonomy in the Age of Knowledge Graphs: Vision and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):13:1–13:22, 2023. doi:10.4230/TGDK.1.1.13.

3 Andrei Ciortea. WebAgents Community Group Report on Interoperability for Agents on the Web. W3C Draft Community Group Report, World Wide Web Consortium, August 29 2025. URL: https://w3c-cg.github.io/webagents/TaskForces/Interoperability/Reports/report-interoperability.html.

4 Andrei Ciortea, Simon Mayer, Fabien Gandon, Olivier Boissier, Alessandro Ricci, and Antoine Zimmermann. A Decade in Hindsight: The Missing Bridge Between Multi-Agent Systems and the World Wide Web. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS'19, Montreal, QC, Canada, May 13-17, 2019*, pages 1659–1663. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

5 Marina De Vos, Sabrina Kirrane, Julian Padget, and Ken Satoh. ODRL Policy Modelling and Compliance Checking. In Paul Fodor, Marco Montali, Diego Calvanese, and Dumitru Roman, editors, *Rules and Reasoning – Third International Joint Conference, RuleML+RR 2019, Bolzano, Italy, September 16-19, 2019, Proceedings*, volume 11784 of *Lecture Notes in Computer Science*, pages 36–51. Springer, September 2019. doi:10.1007/978-3-030-31095-0.

6 Beatriz Esteves, Andrés Chomczyk Penedo, Blessing Mutiro, Haleh Asgarinia, and Dave Lewis. Privacy Paradigm ODRL Profile. Project report, PROTECT Consortium, April 11 2022. URL: https://w3id.org/ppop.

7 Beatriz Esteves, Delaram Golpayegani, Georg P. Krog, Harshvardhan J. Pandit, Julian Flake, and Paul Ryan. Digital Privacy Vocabulary (DPV), version 2.1. Final Community Group Report, World Wide Web Consortium, March 16 2025. URL: https://w3c.github.io/dpv/2.1/dpv/.

8 Beatriz Esteves, Harshvardhan J. Pandit, and Víctor Rodríguez-Doncel. ODRL Profile for Expressing Consent through Granular Access Control Policies in Solid. In *IEEE European Symposium on Security and Privacy Workshops, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*, pages 298–306, 2021. doi:10.1109/EuroSPW54576.2021.00038.

9 Beatriz Esteves and Víctor Rodríguez-Doncel. Analysis of ontologies and policy languages to represent information flows in GDPR. *Semantic Web Journal*, 15(3):709–743, 2024. doi:10.3233/SW-223009.

10 Nicoletta Fornara, Victor Rodríguez-Doncel, Beatriz Esteves, Simon Steyskal, Benedict Whittam Smith, and Yassir Sellami. ODRL Formal Semantics. Draft Community Group Report, World Wide Web Consortium, July 02 2025. URL: https://w3c.github.io/odrl/formal-semantics/.

11 Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. To Be High-Risk, or Not To Be – Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 905–915, New York, NY, USA, June 2023. Association for Computing Machinery. doi:10.1145/3593013.3594050.

12 Delaram Golpayegania, Harshvardhan J. Panditb, Declan O'Sullivana, and Dave Lewisa. Semantic Frameworks to Support Implementation of the EU AI Act. *OSF Preprints*, 2025. Submitted to Computer Law & Security Review. URL: https://doi.org/10.31219/osf.io/43rq2_v1.

13 Arghavan Hosseinzadeh, Robin Brandstädter, and Jessica Chwalek. ODRL Profile for Data Sovereignty. Odrl profile, Fraunhofer IESE, September 9 2024. URL: https://w3id.org/ods/.

14 Renato Ianella. Open Digital Rights Language (ODRL) Version 1.1. W3C Note, World Wide Web Consortium, September 19 2002. URL: http://www.w3.org/TR/2002/NOTE-odrl-20020919/.

15 Renato Ianella and Serena Villata. ODRL Information Model 2.2. W3C Recommendation, World Wide Web Consortium, February 15 2018. URL: http://www.w3.org/TR/2018/REC-odrl-model-20180215/.

16 Renato Iannella, Michael Steidl, Stuart Myles, and Víctor Rodríguez-Doncel. ODRL Vocabulary & Expression 2.2. W3C Recommendation, World Wide Web Consortium, February 15 2018. URL: https://www.w3.org/TR/2018/REC-odrl-vocab-20180215/.

17 IPTC Rights Expression Working Group. IPTC RightsML Standard 2.0. Odrl profile, International Press Telecommunications Council, August 6 2018. URL: https://www.iptc.org/std/RightsML/2.0/RightsML_2.0-specification.html.

18 Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian A. Padget, Terry R. Payne, Munindar P. Singh, Valentina Tamma, and Antoine Zimmermann. Governance of Autonomous Agents on the Web: Challenges and Opportunities. *ACM Transactions on Internet Technologies*, 22(4):104:1–104:31, 2022. doi:10.1145/3507910.

19 Holger Knublauch, Thomas Bergwinkl, Yousouf Taghzouti, and Simon Werner. SHACL 1.2 Core. W3C First Public Working Draft, World Wide Web Consortium, March 18 2025. URL: https://www.w3.org/TR/2025/WD-shacl12-core-20250318/.

20 Holger Knublauch and Dimitris Kontokostas. Shapes Constraint Language (SHACL). W3C Recommendation, World Wide Web Consortium, July 20 2017. URL: https://www.w3.org/TR/2017/REC-shacl-20170720/.

21 Penny Labropoulo and Victor Rodríguez-Doncel. ODRL Profile for Policies of Language Resources and Technologies. W3C Community Group Draft Report, World Wide Web Consortium, June 2 2025. URL: https://rdflicense.linkeddata.es/profile.html#.

22 Mo McRoberts and Victor Rodríguez Doncel. Open Digital Rights Language (ODRL) Ontology. Community Group draft, World Wide Web Consortium, May 12 2014. URL: https://www.w3.org/ns/odrl/2/ODRL20.

23 Open Mobile Alliance. DRM Specification, Approved Version 2.0.2. OMA Technical Specification, Open Mobile Alliance, July 23 2008. URL: https://www.openmobilealliance.org/release/DRM/V2_0_2-20080723-A/OMA-TS-DRM_DRM-V2_0_2-20080723-A.pdf.

24 Harshvardhan J. Pandit and Beatriz Esteves. Enhancing Data Use Ontology (DUO) for Health-Data Sharing by Extending it with ODRL and DPV. *Semantic Web Journal*, 15(4):1473–1498, 2024. doi:10.3233/SW-243583.

25 Harshvardhan J. Pandit, Beatriz Esteves, Georg P. Krog, Paul Ryan, Delaram Golpayegani, and Julian Flake. Data Privacy Vocabulary (DPV) – Version 2.0. In Gianluca Demartini, Katja Hose, Maribel Acosta, Matteo Palmonari, Gong Cheng, Hala Skaf-Molli, Nicolas Ferranti, Daniel Hernández, and Aidan Hogan, editors, *The Semantic Web – ISWC 2024 – 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part III*, volume 15233, pages 171–193, Cham, October 2024. Springer. doi:10.1007/978-3-031-77847-6_10.

26 Siem Velmaat. Automated machine-readable data access agreements by applying ODRL to a FAIR Data Train. Master's thesis, University of Twente, September 2024. URL: https://essay.utwente.nl/103662/1/Veltmaat_MA_EEMCS.pdf.

27 Benedict Whittam Smith and Mark Bird. Market Data Profile for ODRL 1.0. W3C Community Group Draft Report, World Wide Web Consortium, December 1 2021. URL: https://www.w3.org/2021/md-odrl-profile/v1/.

28 Antoine Zimmermann, Andrei Ciortea, Catherine Faron, Eoin O'Neill, and María Poveda-Villalón. Pody: A Solid-Based Approach to Embody Agents in Web-Based Multi-Agent-Systems. In Andrei Ciortea, Mehdi Dastani, and Jieting Luo, editors, *Engineering Multi-Agent Systems – 11th International Workshop, EMAS 2023, London, UK, May 29-30, 2023, Revised Selected Papers*, volume 14378 of *Lecture Notes in Computer Science*, pages 220–229. Springer, 2023. doi:10.1007/978-3-031-48539-8_15.

Participants

- Nirav Ajmeri
University of Bristol, GB
- Matthew Arrott
Coactive Computing, US
- Matteo Baldoni
University of Turin, IT
- Cristina Baroglio
University of Turin, IT
- Victor Charpenay
Mines Saint-Étienne, FR
- Amit K. Chopra
Lancaster University, GB
- Mehdi Dastani
Utrecht University, NL
- Marina De Vos
University of Bath, GB
- Davide Dell'Anna
Utrecht University, NL
- Frank Dignum
University of Umeå, SE
- Beatriz Esteves
Ghent University, BE
- Nicoletta Fornara
USI – Lugano, CH
- Joris Hulstijn
Utrecht University, NL
- Timotheus Kampik
SAP Berlin, DE &
Umeå University, SE
- Nadin Kokciyan
University of Edinburgh, GB
- Beishui Liao
Zhejiang University, CN
- Réka Markovich
University of Luxembourg, LU
- Pradeep Murukannaiah
TU Delft, NL
- Vivek Nallur
University College Dublin, IE
- Luis Gustavo Nardin
IMT Mines Saint-Étienne, FR
- Sebastian Neumaier
FH – St. Pölten, AT
- Julian Padget
University of Bath, GB
- Victor Rodriguez Doncel
Polytechnic University of
Madrid, ES
- Susana Rodríguez Verdugo
Kunveno Digital – Madrid, ES
- Ken Satoh
Research Organization of
Information and Systems –
Tokyo, JP
- Jaime Sichman
University São Paulo, BR
- Judith Simon
Universität Hamburg, DE
- Munindar P. Singh
North Carolina State University –
Raleigh, US
- Simon Steyskal
Siemens AG – Wien, AT
- Sz-Ting Tzeng
University of Umeå, SE
- Leon van der Torre
University of Luxembourg, LU
- Harko Verhagen
Stockholm University, SE
- Rigo Wenning
W3C / ERCIM, FR
- Jessica Woodgate
University of Bristol, GB
- Pinar Yolum
Utrecht University, NL
- Antoine Zimmermann
Ecole des Mines –
St. Étienne, FR

Report from Dagstuhl Seminar 25272

Challenges of Human Oversight: Achieving Human Control of AI-Based Systems

Markus Langer*[1], Raimund Dachselt*[2], Q. Vera Liao*[3], Tim Miller*[4], and Nava Tintarev*[5]

1 **Universität Freiburg, DE.** markus.langer@psychologie.uni-freiburg.de
2 **TU Dresden, DE.** raimund.dachselt@tu-dresden.de
3 **Microsoft – Montréal, CA.** veraliao@umich.edu
4 **University of Queensland – Brisbane, AU.** timothy.miller@uq.edu.au
5 **Maastricht University, NL.** n.tintarev@maastrichtuniversity.nl

Abstract

Human oversight is a key safeguard for AI systems, intended to mitigate risks by adding a human layer of safety and control. Oversight personnel should, for example, detect malfunctions or violations of fundamental rights such as discriminatory decision-making and intervene accordingly. Human oversight is also central to AI governance and ethics, and is mandated by Articles 14 and 26 of the EU AI Act for high-risk AI. This Dagstuhl Seminar brought together experts from artificial intelligence, human-computer interaction, human factors and psychology, philosophy and ethics, and law to explore conceptual, technical, legal, and practical dimensions of human oversight of AI. Across the seminar, participants provided perspective talks from the different disciplines and engaged in working groups and use-case specific discussions in order to establish a science of human oversight of AI systems. The main outcome of this seminar is a general framework that outlines the architecture, processes, and sociotechnical design dimensions of human oversight of AI systems.

Seminar June 29 – July 4, 2025 – https://www.dagstuhl.de/25272
2012 ACM Subject Classification Human-centered computing → HCI design and evaluation methods; Human-centered computing → HCI theory, concepts and models
Keywords and phrases artifical intelligence, explainable ai, human oversight, norms and regulations, safety
Digital Object Identifier 10.4230/DagRep.15.6.189

1 Executive Summary

Markus Langer (Universität Freiburg, DE)
Raimund Dachselt (TU Dresden, DE)
Q. Vera Liao (Microsoft – Montréal, CA)
Tim Miller (University of Queensland – Brisbane, AU)
Nava Tintarev (Maastricht University, NL)

What is effective human oversight of AI systems? The Dagstuhl Seminar 25272 "Challenges of Human Oversight: Achieving Human Control of AI-Based Systems" brought together interdisciplinary experts from artificial intelligence, human-computer interaction, human factors and psychology, philosophy and ethics, as well as law to explore conceptual, technical,

* Editor / Organizer

Challenges of Human Oversight: Achieving Human Control of AI-Based Systems, *Dagstuhl Reports*, Vol. 15, Issue 6, pp. 189–204
Editors: Markus Langer, Raimund Dachselt, Q. Vera Liao, Tim Miller, and Nava Tintarev
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

legal, and practical dimensions of human oversight of AI. Across the seminar, participants provided perspective talks from the different disciplines and engaged in working groups and use-case specific discussions in order to establish a science of human oversight of AI systems. The main outcome of this seminar is a general framework that outlines the architecture, processes, and sociotechnical design dimensions of human oversight of AI systems. In the following, we present some of the key insights of this seminar in more detail.

Conceptual Foundations of Human Oversight. Human oversight is defined as a human activity to monitor and intervene in AI-supported tasks (typically at runtime) with the aim of sufficiently mitigating risks. Mitigating risks means detecting errors, system malfunctions, or inadequate outputs. Effectiveness depends on epistemic access, causal power, self-control, and fitting intentions of the human oversight personnel. In order to optimize the human oversight effectiveness, it requires designing the sociotechnical dimensions of human oversight: the human factors, technical design, and contextual considerations. Human oversight can operate at multiple layers and across distributed roles within human oversight teams and is inherently interdependent with other risk mitigation measures.

Human Factors, Technical Design, and Contextual Considerations. Human factors cover situation awareness, decision making, cognitive biases, workload, motivation, training, and collaboration between oversight personnel. Technical design must support human detection of system errors and failures, for example via visualization, technical support tools, adaptive automation, handover design, and personalization. Contextual considerations to support human oversight include time and resource requirements for effective human oversight as well as the clarity of human oversight roles and duties.

Legal, and Normative Considerations. Human oversight effectiveness requires normative judgment beyond legal mandates. Human oversight objects include individual AI systems, highly-autonomous agentic AI in high-risk domains, as well as AI systems operated by users such as patients using mental health chatbots who themselves are not considered human oversight personnel. The seminar highlighted the importance to consider the relation between human oversight, other risk management measures, and technical standards.

Evaluating Human Oversight Effectiveness. Evaluation of human oversight implementation is crucial given that human oversight effectiveness can only be achieved iteratively. Metrics include effectiveness of monitoring and interventions (e.g., detecting erroneous AI outputs, overriding these outputs), alignment with human oversight protocols, and long-term performance outcomes. Mixed-methods approaches (quantitative and qualitative) and comparative studies of different human oversight design options (e.g., varying support interfaces) were discussed as possible options to evaluate human oversight effectiveness.

Human Oversight Effectiveness as an Iterative and Multi-Layered Challenge. Continuous updating of human oversight design is essential, integrating empirical feedback and ensuring institutional support for high-quality and sustainable human oversight of AI. Furthermore, we saw that effective oversight required identifying information and workflows across regulatory, technical, and interface layers.

Conclusion. The seminar demonstrated that human oversight of AI is a multifaceted, interdisciplinary challenge, involving conceptual clarity, human factors, technical design, contextual considerations, evaluation frameworks, as well as legal and ethical considerations. The outputs of this seminar provide a foundation for theoretical modeling, empirical research, practical design guidance, and normative reflection, establishing a roadmap for advancing

effective human oversight in AI systems contributing to the safe implementation of AI in high-risk contexts. Next steps include joint publications (e.g., a framework for human oversight of AI), developing technical support tools for effective human oversight, and community building through workshops at key human-computer interaction and AI conferences.

2 Table of Contents

3 Overview of Talks

Before the seminar, we prompted input on the topic of human oversight of AI from the different disciplines that were part of the seminar. We selected five talks to be presented during Monday and Tuesday of the seminar that provided a starting point for interdisciplinary discussions on human oversight of AI.

3.1 What is Human Oversight?

Markus Langer (Universität Freiburg, DE) and Sarah Sterz (Universität des Saarlandes – Saarbrücken, DE)

Joint work of Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, Markus Langer
Main reference Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, Markus Langer: "On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives", in Proc. of the The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024, pp. 2495–2507, ACM, 2024.
URL https://doi.org//10.1145/3630106.3659051

Human oversight is discussed as a potential safeguard to mitigate risks in AI applications. This prompts a critical examination of the role and conditions necessary for effective or meaningful human oversight of these systems. Based on the claim that the main objective of human oversight is risk mitigation, we propose a viable understanding of effectiveness in human oversight: for human oversight to be effective, the oversight person has to have (a) sufficient control and power with regard to the system and its effects, (b) suitable epistemic access to relevant aspects of the situation, and (c) fitting intentions for their role. Furthermore, we argue that this is equivalent to saying that an oversight person is effective if and only if they are morally responsible and have fitting intentions. Against this backdrop, we suggest facilitators and inhibitors of effectiveness in human oversight when striving for practical applicability. We discuss factors in three domains, namely, the technical design of the system, individual factors of oversight persons, and the environmental circumstances in which they operate.

3.2 Human Oversight of AI Systems: An HCI Perspective

Ujwal Gadiraju (TU Delft, NL)

The systematic involvement of humans in monitoring, controlling, and intervening in AI system operations to ensure safety, accountability, and alignment with human values has a central role to play in regulating meaningful AI adoption. This talk synthesizes challenges and opportunities for human oversight of AI systems from a human-computer interaction (HCI) standpoint by addressing the overarching goal of designing and developing interfaces, interaction paradigms, and workflows that enable effective human-AI collaboration while maintaining meaningful human control. This perspective talk concludes with a discussion around the critical components of human oversight, and potential methods and measures for effective human oversight.

3.3 A Legal Perspective on Human Oversight

Anne Lauber-Rönsberg (TU Dresden, DE)

Objectives that can be pursued by human oversight are risk mitigation, trustworthiness of AI systems, human agency and autonomy, human-centered AI and accountability. There are many facets of human involvement that contribute to these objectives: human approval of AI output (human in the loop); monitoring of AI decision processes and the possibility to intervene (human on the loop); reducing the role of the AI-system to supporting human decisions; human interventions in the operation of the AI system, such as correction of inaccurate results or the infamous stop button and finally a human examination of AI-output after contestation. These activities are either executed during run time, maybe even real time, during inspection time and supported by decisions on the technical design during design time. Monitoring activities can either relate to every output or to selected samples. They may be always required, for instance by the law, or only be triggered upon the request of an affected person. I highlight that the concept of human oversight of Art. 14 and 26 of the EU AI Act is much more confined since it relates only to a subset of the activities named. Human oversight for high-risk AI systems under the AI Act aims at mitigating risks to health, safety or fundamental rights and must be be commensurate with the risks, level of autonomy and context of use, has to take place during the use of the system and requires the ability of the oversight person to intervene in the operation of the AI system, such as by stopping it or by overriding its output. Apart from this, the AI Act contains few specifications regarding technical design. Thus, the main responsibility is on the provider (= developer), who has to determine appropriate oversight measures that are either built into the system (such as human-machine interface tools or operational constraints) or have to be executed by the deployer in line with the instructions for use. The deployer has to assign human oversight to persons who have the necessary competence, training and authority and needs to ensure the necessary support. Oversight persons can be employees who use the AI system for their work or external third parties. End-users in a private context are not required to conduct human oversight by the AI Act. Questions to be discussed are: Under which conditions and in which contexts can human oversight be regarded as (sufficiently) effective? How can technical standards be drafted to clarify the legal obligations? How can human oversight be implemented in case of more autonomous agentic AI systems? What applies in situations where there is no deployer (e.g., in case of smart toys)? The legal obligations to provide human oversight are not applicable to AI systems that are only "intended to perform a preparatory task". How is this exception to be construed?

3.4 Towards Human Oversight of Imperfect Automation: A Dagstuhl CELLAR (Cognitive Engineering Lessons Learned And Reflections) Perspective

Tim Miller (University of Queensland – Brisbane, AU) and Liz Sonenberg (University of Melbourne, AU)

In this perspective talk, we profile a selection of research from cognitive science and cognitive systems engineering that has high relevance to human oversight in situations involving artificial intelligence. We look back at research on human interaction with imperfect automation, that captures automation-induced failures that can occur at a technical level, through the human-automation interaction, and/or through broader sociotechnical systems breakdowns. We show there is a rich existing literature that can inform contemporary analyses of human oversight of AI systems, and illustrate that many of the "new" problems in human oversight of AI are not so new at all. We also present a more future-oriented view of emerging paradigms in cognitive science/psychology, and how these may affect approaches to characterizing and designing the AI and the great sociotechnical system in which it operates.

3.5 From Traditional Auditing to Everyday Oversight: The Role of Users in Algorithmic Accountability

Motahhare Eslami (Carnegie Mellon University - Pittsburgh, US)

This talk examines a growing shift in algorithmic accountability: from formal, expert-led audits to more informal, everyday forms of oversight performed by users themselves. While traditional audits remain essential, they often fail to capture the full range of harms experienced by the public – particularly those affecting marginalized communities. In contrast, users are increasingly taking initiative to investigate, document, and call out algorithmic failures, often through grassroots experimentation and collaborative sensemaking.

Through a series of empirical studies, the talk traces how users have identified and investigated algorithmic harms across platforms. These forms of "everyday oversight" not only surface harms often missed by formal audits but also broaden the notion of who gets to hold AI systems accountable. To support this growing public role, the talk introduces WeAudit, a platform co-designed with users and industry practitioners to facilitate participatory auditing at scale. By recognizing users as legitimate and capable auditors, WeAudit aims to institutionalize support for community-driven accountability efforts.

4 Working groups

From Tuesday to Thursday, main activities were organized in working groups where we worked on conceptual and design challenges of human oversight of AI. In the following, we provide summaries of the working groups. The order of authors in the working groups is alphabetical.

4.1 Thematic Analysis of Lightning Talks

Raimund Dachselt (TU Dresden, DE), Susanne Gaube (University College London, GB), Tim Miller (University of Queensland – Brisbane, AU), Liz Sonenberg (University of Melbourne, AU), and Nava Tintarev (Maastricht University, NL)

This working groups conducted a thematic analysis of the lightning talks, where participants introduced their research and discussion interests during the first day of the seminar. The analysis surfaced a rich landscape of perspectives on human oversight, which clustered into four broad thematic areas. First, participants highlighted applications and use cases, supported by a taxonomy of types of human oversight and intervention, and situated oversight in relation to adjacent concepts such as accountability and responsibility. Second, a legal, risk, and ethics theme emerged, including references to the EU AI Act, legal processes, risk management practices, and the need for future-proof regulatory frameworks. Third, participants stressed the role of human factors, from cognitive limitations and strengths to capability development, user interaction design, and explainable AI. Finally, a set of themes focused on design, evaluation, and technical development, encompassing design methodologies, information visualization, evaluation and measurement approaches, traceability, modeling, and the technical capabilities of AI systems. Together, these clusters reflect the interdisciplinary scope of the seminar and lay a foundation for deeper discussions on human oversight.

4.2 Black Mirror Writers' Room Exercise

Nava Tintarev (Maastricht University, NL)

Authors for this working group are all participants. Nava Tintarev moderated the working group.

Black Mirror is science-fiction series that offers a critical discussion and implications of technological developments that are expected in the near future (5 years). Each episode focuses on a specific technology, e.g., social credit scores in the episode "Nosedive", inspired by the social credit system developed in China. All of these episodes are dystopian, looking at the dark side of technology (dystopian means relating to or denoting an imagined state or society where there is great suffering or injustice). The task for participants was to write their own Black Mirror scenario involving a specific task context relevant to the topic of human oversight of AI. These are the teasers of the Black Mirror episodes pitched by the participant groups.

Train Wreck. An employee of the government transport department notices massive chaos in public transport. When questioning his superiors, he finds out that the cause is decisions made by his Gov3.0 productive tool, that he was incentivized to use for efficiency reasons... and that he authorized himself.

Border Control AI. Otto once found deep meaning in his work as a border agent, making autonomous decisions that impacted national security – but as AI systems take over, his role becomes reduced to a mere formality. Despite new oversight measures meant to ensure human responsibility, Otto feels powerless, resorting to workarounds while struggling with burnout, dwindling job prospects, and a growing sense of despair.

AI Teacher. A suicide epidemic sweeps the country. Unemployment is high and there is an overabundance of unhappy bus drivers. In the meantime care jobs are left unfilled. Jeremy the high-school student wants to study nursing, but his personalized AI tutor only teaches traffic regulations and driving simulations. His parents ask the AI tutor for information about other personalized curricula. While the parents do not get any answers, this starts a positive cascade to the district educational office and the AI allocation module. Follow these parents in their brave battle against bureaucratic demons.

Professor Bot. A university introduces a new software for professors to create digital copies of themselves to "scale" advising. Professor Best had the best intention by creating a "better" version of his digital copy ... Who will stay? Professor Best or ProfessorBot?

BlueSky. The world is in an energy crisis. We have exhausted oil and making no progress in fusion. The governments have opened a Blue Sky grant call for groundbreaking ideas on how to solve the energy crises. Generation one of grants has offered no impact. A generation two call is out. Alex is considering a proposal. They found a old fashioned organic book in their basement on particle physics. Alex makes a proposal but it is rejected. Forty years ago research peer review was replaced by AI. Particle physics has been systematically denied funding because it is very low practical impact and there are real problems in society.

4.3 Use Case Human Resource Management

Anna Maria Feit (Universität des Saarlandes – Saarbrücken, DE), Harmanpreet Kaur (University of Minnesota – Minneapolis, US), Mark T. Keane (University College Dublin, IE), Richard Landers (University of Minnesota – Minneapolis, US), Markus Langer (Universität Freiburg, DE), and Q. Vera Liao (Microsoft – Montréal, CA)

This working group explored human oversight in job–applicant matching platforms, framing it as a process of detecting failures and taking appropriate action. Oversight applies at two levels: monitoring algorithmic performance and reviewing individual matches. The group distinguished between aggregate risks (e.g., long-term discrimination) and individual risks. Detection of inadequate system outputs may rely on automated thresholds or manual checks, while interventions range from alerting decision makers and redoing tasks to retraining models or redesigning the system.

Effective oversight requires dashboards and feedback loops that integrate aggregate and individual data, reliable alerts, and clarity about the authority of oversight roles. Challenges include avoiding overreliance on oversight, balancing efficiency with risk mitigation, and defining what a "safe state" means in non-real-time systems. At a meta-level, oversight design must be continuously updated as new problems arise, with institutions ensuring its quality over time.

This working group also discussed human oversight of coaching AI, where a human oversight person oversees an interaction between a conversational AI and a user. This use case also inspired discussions on other high-risk conversational AI domains such as AI in psychotherapy, where human oversight may require a control room where human oversight personnel can oversee and intervene in patient-AI conversations, or AI used by children where human oversight may be performed by parents.

4.4 Use Case Autonomous Driving

Oana Inel (Universität Zürich, CH), Linda Onnasch (TU Berlin, DE), and Carola Plesch (BSI – Bonn, DE)

This working group explored human oversight in the use case of autonomous driving. For example, a remote oversight person oversees a fleet of vehicles within a defined area, but without continuous monitoring. Instead, autonomous vehicles escalate critical situations to the oversight person, while time-critical decisions remain with the vehicle itself. Passengers have no direct intervention options beyond contacting the oversight person.

The group characterized this setting as remote, multitasking oversight, with oversight tasks interrupting other activities through forced task switches. Key questions included whether a "safe state" exists for the vehicle, and whether it can be unconditionally enforced. Intervention options available to the oversight person largely concern planning decisions (e.g., route selection), escalation measures (e.g., contacting a task force), or putting a vehicle into a safe state. Interfaces must integrate multiple streams of information – traffic, weather, vehicle status – while supporting interaction with passengers and prioritization across simultaneous requests.

The group identified significant challenges for oversight effectiveness, such as establishing situation awareness from a distance, coping with limited multimodal information, managing authority boundaries between human and automation, and handling fluctuating workloads. For effective human oversight, the group emphasized the importance of contextual information provided prior to takeover, such as vehicle type, location, reason for the request, and possible courses of action, potentially offered through structured, pre-selected options.

4.5 Defining and Conceptualizing Human Oversight

Kevin Baum (DFKI – Saarbrücken, DE), Markus Langer (Universität Freiburg, DE), Anne Lauber-Rönsberg (TU Dresden, DE), Johann Laux (University of Oxford, GB), Tim Schrills (Universität Lübeck, DE), and Sarah Sterz (Universität des Saarlandes – Saarbrücken, DE)

This working group focused on defining and conceptualizing human oversight of AI systems. They proposed that human oversight consists of monitoring and intervening in AI-supported tasks with the explicit aim of mitigating risks. Oversight is effective only when these risks are sufficiently reduced. The group emphasized that effectiveness depends not only on individual abilities and motivation, but also on technical design, organizational context, and available interventions.

They identified four key factors – epistemic access, causal power, self-control and fitting intentions – that shape oversight effectiveness, and noted that different layers (legal, institutional, design/technical) influence these factors in distinct ways. Human oversight was seen as the last layer of risk mitigation, highly interdependent with other safeguards, and potentially distributed across multiple people and roles.

The group stressed that while training, design, and organizational conditions are not themselves "oversight," they are necessary enablers for it to be effective. A structured overview of oversight activities and effectiveness factors can serve as a foundation for theoretical models and guide empirical research.

4.6 Dimensions of Human Oversight of AI

Virginia Dignum (University of Umeå, SE), Ujwal Gadiraju (TU Delft, NL), Brian Lim (National University of Singapore, SG), Marija Slavkovik (University of Bergen, NO), Chenhao Tan (University of Chicago, US), Ziang Xiao (Johns Hopkins University – Baltimore, US), and Hanwei Zhang (Universität des Saarlandes – Saarbrücken, DE)

This working group outlined an dimensions of human oversight of AI. These dimensions included oversight goals, oversight tasks, oversight persons and their characteristics and skills, oversight failures and oversight evaluation. This working group also discussed the importance of proportionality of human oversight depending on the context of use and the efficient design of interfaces to support human oversight.

4.7 Challenges of Human Oversight – Human Factors Challenges

Anna Maria Feit (Universität des Saarlandes – Saarbrücken, DE), Liz Sonenberg (University of Melbourne, AU), Markus Langer (Universität Freiburg, DE), Q. Vera Liao (Microsoft – Montréal, CA), and Linda Onnasch (TU Berlin, DE)

This working group focused on the human factors and design considerations of oversight systems. For run-time oversight, they highlighted classic human factors issues such as information processing, decision making, cognitive and automation biases, uncertainty, attention management, workload, motivation, training, and collaboration. This was also described as "old human factors wine in a new bottle."

They also discussed design requirements for technical components that support human oversight: interfaces and alerts, signal processing and visualization, tools for decision support and intervention, transparency of both AI and the broader human–AI task system, history tracking, partial automation of monitoring or actions, handovers, and personalization for oversight personnel.

Finally, the group noted the dependencies between oversight systems and the task systems they oversee, and raised questions about human factors challenges for non–run-time overseers (e.g., auditors), suggesting that interface and tool design should differ across oversight roles.

4.8 Challenges of Human Oversight – Technical Challenges

Brian Lim (National University of Singapore, SG), Chenhao Tan (University of Chicago, US), Ziang Xiao (Johns Hopkins University – Baltimore, US), and Hanwei Zhang (Universität des Saarlandes – Saarbrücken, DE)

This working group investigated the technical challenges of enabling and supporting human oversight. They structured their discussion around four main dimensions. First, preparing information for oversight, including explainability, interpretability, calibration, and aggregation of explanations. Second, monitoring, with emphasis on detecting issues through structured information flows, context operationalization, robustness, adaptivity to drift, and third-party risk assessments. Third, intervention, covering controllability and steerability of AI models, identifying and operationalizing fail-safe states, fallback and recovery mechanisms, and managing hyperparameter sensitivity. Fourth, testing and evaluation, highlighting the value of sandbox environments, simulation, data synthesis, long-term evaluation, and methods for quantifying both human effort and systemic risks.

Additional themes included the need for personalization versus generalization in oversight tools, ensuring scalability, efficiency, and privacy (e.g., double-blind oversight), and recognizing impossibility results in operationalizing oversight (as with fairness definitions). The group underscored the importance of addressing these challenges in a cost-effective way while safeguarding both human and system effectiveness.

4.9 Challenges of Human Oversight – Legal Challenges

Anne Lauber-Rönsberg (TU Dresden, DE), Johann Laux (University of Oxford, GB), Philip Meinel (TU Dresden, DE), and Silja Voeneky (Universität Freiburg, DE)

This working group examined normative, legal, and organizational dimensions of effective human oversight. They noted that while effectiveness cannot be determined by legal criteria alone, any threshold of "sufficient" oversight requires normative judgment. Discussion centered on the objects of oversight, such as which AI systems and users of AI should be overseen, including highly autonomous or composite systems in high-risk domains. This raised questions about whether the EU AI Act permits or restricts highly autonomous AI in such contexts.

The group also considered the relationship between human oversight and other risk mitigation measures, identifying potential sources of failure in relation to fundamental rights and ways oversight might address them. They highlighted the importance of role distribution, including centralized versus distributed oversight, requirements for outsourcing oversight functions, and the role of deployers in supporting effective oversight. Finally, the group noted the need for technical standards to guide the design and implementation of oversight in practice.

4.10 Challenges of Human Oversight – Evaluation Challenges

Raimund Dachselt (TU Dresden, DE), Susanne Gaube (University College London, GB), Holger Hermanns (Universität des Saarlandes – Saarbrücken, DE), Oana Inel (Universität Zürich, CH), Mark T. Keane (University College Dublin, IE), Tim Miller (University of Queensland – Brisbane, AU), Carola Plesch (BSI – Bonn, DE), and Nava Tintarev (Maastricht University, NL)

This working group focused on evaluation approaches for human oversight. They emphasized the value of both comparative studies (e.g., different interface designs) and mixed methods combining quantitative with qualitative analysis of human oversight effectiveness. The group proposed metrics at multiple levels: organizational (e.g., documentation, global success indicators, costs); oversight personnel (e.g., knowledge about the domain/AI model/oversight task, motivation, efficiency, cognitive load); AI systems (e.g., performance with and without oversight); and the oversight support technology (e.g., interpretability, effectiveness in detection, predictive performance). For interventions, they suggested assessing both the effectiveness and efficiency of actions (e.g., time to resolution, coverage of delegation, alignment with protocols). Together, these dimensions offer a comprehensive framework for evaluating oversight across technical, human, and organizational layers.

5 Integration and Outlook

On Thursday evening and Friday, the seminar conducted integrative working group and a collective outlook session. In the following, we summarize these sessions.

5.1 Bringing it all Together

Raimund Dachselt (TU Dresden, DE), Markus Langer (Universität Freiburg, DE), Q. Vera Liao (Microsoft – Montréal, CA), Tim Miller (University of Queensland – Brisbane, AU), and Nava Tintarev (Maastricht University, NL)

All participants were part of this session. It was moderated by the organizers of the seminar.

This working group led to the development of a general human oversight architecture, including the AI task to be overseen, the human oversight task, the human oversight personnel, hierarchical layers of human oversight, and sociotechnical human oversight support design. It also led to the definition of human oversight processes and tasks, including monitoring and intervention in AI operations. Additionally, this working group synthesized challenges of human oversight in practice, such as the context-dependency and effectiveness of human oversight, challenges in evaluating human oversight, and challenges in complying with human oversight regulations.

5.2 Future Activities

Raimund Dachselt (TU Dresden, DE), Markus Langer (Universität Freiburg, DE), Q. Vera Liao (Microsoft – Montréal, CA), Tim Miller (University of Queensland – Brisbane, AU), and Nava Tintarev (Maastricht University, NL)

During the final day, the entire group discussed next steps and opportunities for collaboration. Several joint publications were proposed, including a framework paper on human oversight, a taxonomy of scenario attributes and oversight requirements, a design fiction paper based on the "Black Mirror" exercise, and follow-up work on the overlay of the EU AI Act and sustainable human oversight. Further ideas included papers on legal obligations around high-risk AI systems, dual-use risks of oversight, and a possible Dagstuhl Manifesto synthesizing outcomes.

The group also identified venues for dissemination and engagement, such as CHI, FAccT, IUI, CSCW and AI-focused conferences (NeurIPS, ICML, AAAI). Planned activities include workshops at major venues, exploration of special issues in journals (e.g., AI Magazine, Technology, Mind and Behavior), and the development of a software toolkit to make oversight testable.

Finally, the group outlined opportunities for funding and collaboration, including a COST Action proposal, EU and Australian grants, and partnerships such as the Dutch National Police project. They also encouraged exchanges, such as PhD visits across participating institutions, to sustain the momentum of the seminar.

Participants

- Kevin Baum
DFKI – Saarbrücken, DE
- Raimund Dachselt
TU Dresden, DE
- Virginia Dignum
University of Umeå, SE
- Anna Maria Feit
Universität des Saarlandes – Saarbrücken, DE
- Ujwal Gadiraju
TU Delft, NL
- Susanne Gaube
University College London, GB
- Holger Hermanns
Universität des Saarlandes – Saarbrücken, DE
- Oana Inel
Universität Zürich, CH
- Harmanpreet Kaur
University of Minnesota – Minneapolis, US
- Mark T. Keane
University College Dublin, IE
- Richard Landers
University of Minnesota – Minneapolis, US
- Markus Langer
Universität Freiburg, DE
- Anne Lauber-Rönsberg
TU Dresden, DE
- Johann Laux
University of Oxford, GB
- Q. Vera Liao
Microsoft – Montréal, CA
- Brian Lim
National University of Singapore, SG
- Philip Meinel
TU Dresden, DE
- Tim Miller
University of Queensland – Brisbane, AU
- Linda Onnasch
TU Berlin, DE
- Carola Plesch
BSI – Bonn, DE
- Tim Schrills
Universität Lübeck, DE
- Marija Slavkovik
University of Bergen, NO
- Liz Sonenberg
University of Melbourne, AU
- Sarah Sterz
Universität des Saarlandes – Saarbrücken, DE
- Chenhao Tan
University of Chicago, US
- Nava Tintarev
Maastricht University, NL
- Silja Voeneky
Universität Freiburg, DE
- Ziang Xiao
Johns Hopkins University – Baltimore, US
- Hanwei Zhang
Universität des Saarlandes – Saarbrücken, DE