DAGSTUHL
REPORTS

Volume 15, Issue 1, January 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
October, 2025

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

Digital Object Identifier: 10.4230/DagRep.15.1.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 25021

Grand Challenges for Research on Privacy Documents

Florian Schaub[*1], Christine Utz[*2], Shomir Wilson[*3], and Lu Xian[†4]

1 **University of Michigan – Ann Arbor, US.** fschaub@umich.edu
2 **Radboud University Nijmegen, NL.** christine.utz@ru.nl
3 **Pennsylvania State University – University Park, US.** shomir@psu.edu
4 **University of Michigan – Ann Arbor, US.** xianl@umich.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25021 "Grand Challenges for Research on Privacy Documents" held in January 2025. This Dagstuhl Seminar gathered an interdisciplinary group of researchers from privacy, natural language processing, human-computer interaction, public policy, and law to identify and characterize key challenges to research on privacy documents, such as privacy policies, terms of use, cookie policies, and other texts about data practices.

Seminar participants worked together to identify and characterize key challenges in privacy document research with the goal of producing a research roadmap for tackling these challenges. Through a series of perspectives talks and panel discussions, participants exchanged experiences in working with privacy documents in research and learned about associated challenges, as well as interdisciplinary intersections and policy considerations. Through deeper engagement in working groups, participants deeply explored research challenges and research directions across five interconnected topics: (1) formats and standardization; (2) datasets, automation, and analysis methods; (3) usable and useful notice and consent; (4) consumer privacy beyond notice and choice; and (5) cross-stakeholder engagement.

Seminar January 5–10, 2025 – https://www.dagstuhl.de/25021
2012 ACM Subject Classification Applied computing → Law, social and behavioral sciences; Computing methodologies → Natural language processing; Security and privacy → Human and societal aspects of security and privacy; Social and professional topics → Privacy policies
Keywords and phrases Human-Computer Interaction, Machine Learning, Natural Language Processing, Privacy Policy, Public Policy
Digital Object Identifier 10.4230/DagRep.15.1.1
Funding Organizers Schaub and Wilson would like to acknowledge support by the National Science Foundation under Award No. 2105734 and 2105736 ("Collaborative Research: SaTC: CORE: Medium: A Large-Scale, Longitudinal Resource to Advance Technical and Legal Understanding of Textual Privacy Information").

1 Executive Summary

Shomir Wilson (Pennsylvania State University – University Park, US)
Florian Schaub (University of Michigan – Ann Arbor, US)
Christine Utz (Radboud University Nijmegen, NL)

The five-day Dagstuhl Seminar "Grand Challenges for Research on Privacy Documents" gathered an interdisciplinary group of researchers from privacy, natural language processing, human-computer interaction, public policy, and law to identify and characterize key challenges to research on privacy documents, such as privacy policies, terms of use, cookie policies,

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

and other texts about data practices. In the status quo, privacy documents primarily serve the compliance needs of companies, while failing to fulfill the needs of other stakeholders in our information society. Although many Internet users have concerns about their privacy, most lack the time, knowledge, and other resources to understand these documents, leaving them underinformed and compromising the goals of the notice and choice paradigm. The needs of other stakeholders, including regulators, researchers, policymakers, and privacy practitioners, are similarly stymied. Although a growing body of research is devoted to analyzing, reconstituting, or otherwise using these documents to satisfy stakeholders' needs, broader interdisciplinary efforts are needed.

The goal of this seminar was to identify and characterize key challenges in privacy document research and to produce a research roadmap of how to tackle them in order to move the field forward. At a high level, the seminar schedule was structured into two stages to produce those outcomes. The first stage consisted of a series of perspectives talks providing background and introductions to relevant disciplines and approaches, as well as thematic panel discussions among participants. In the second stage, participants organized in topical working groups to more deeply explore specific areas and develop elements of the roadmap. Working groups focused on the following themes: (1) document formats and standardization; (2) datasets, automation, and analysis methods; (3) usable and useful notice and consent; (4) consumer privacy beyond notice and choice; and (5) cross-stakeholder engagement.

This report collects abstracts of the three perspective talks and presents research challenges and directions identified in the working groups. Each section describes respective challenges, key research questions, and solution ideas and directions. We present this report to the research community as a resource for discussion and inspiration for future work.

2 Table of Contents

3 Overview of Perspective Talks

3.1 Legal Perspectives on Privacy Documents

Kirsten Martin (Carnegie Mellon University – Pittsburgh, US)

This perspective talk explained why privacy documents have evolved as large, ambiguous documents that do not serve the needs of stakeholders. Privacy laws can be seen as entering a second phase. Where a first phase of privacy regulation focused on the handoff of data to firms as if privacy is relinquished when shared with firms. This placed a heavy burden on privacy notices to ensure individuals choose firms correctly to 'give up' their data. The second phase correctly pivots to recognize that people have privacy interests and rights even when they have shared data with firms.

3.2 NLP/ML Perspectives on Privacy Documents

Sepideh Ghanavati (University of Maine, US)

Privacy policy documents aim to inform users about how their personal information is collected, used, processed, or shared. However, the documents are generally long, contain legal jargon, and include vagueness and ambiguities, which make it hard for the end-users to understand them and make informed decisions. In the last 20 years, researchers leveraged machine learning (ML) and natural language processing (NLP) techniques to create datasets of annotated policies, extract features from the privacy policies, analyze data practices, assess the readability, usability, and utility of privacy policies, and evaluate the consistency and compliance with laws, regulations, and best practices. This talk provided an overview of the state-of-the-art research regarding privacy document analysis with ML/NLP techniques, identifying the key features and contributions, and then provided guidelines and potential research directions for future work.

3.3 Human-centered Perspectives on Privacy Documents

Simone Fischer-Hübner (Karlstad University, Chalmers University of Technology and Gothenburg University, SE)

Privacy by Design can only be achieved if transparency and control functions for users are usable. Therefore, the GDPR is also requiring that data subject rights functionality must be provided in "concise, transparent, intelligible and easily accessible form, using clear and plain language."

This talk first addressed and discussed challenges for human-centered privacy related to privacy documents. These include challenges of notice and consent, the challenge that privacy is only a secondary goal for users, the users' limited rationality and psychological

effects that need to be considered, the challenge that there are no one-size fits all solutions for different types of users and contexts, as well as challenges of explaining privacy-enhancing technologies (PETs) if mentioned in privacy documents.

Secondly, solutions for approaching these challenges and remaining challenges were outlined. It was highlighted that privacy by design of privacy documents needs to be aligned and combined with human-centered and inclusive design. Moreover, various approaches for designing usable privacy notices and usable explanations for PETs were discussed as well as approaches and guidelines for raising the user's attention to essential policy information by engaging them with interactive policy content.

4 Working Groups

4.1 Formats and Standardization

Christine Utz (Radboud University Nijmegen, NL), Rinku Dewri (University of Denver, US), Emma Tosch (Northeastern University – Boston, US), and Lu Xian (University of Michigan – Ann Arbor, US)

Certain firms are legally required to provide privacy documents that notify consumers or end-users of how their data will be collected, used, stored, and transmitted. When firms operate in multiple jurisdictions, they may be required to provide multiple types of privacy documents, each having different information. Furthermore, some firms may wish to provide users with privacy-related information beyond legal requirements. As a result, any discussion of privacy documents involves a many-to-many relationship between heterogeneous stakeholders. This entails studying privacy policies across a variety of sources.

Locating and understanding these different sources of privacy documents in the field is critically important for researchers and legislators performing privacy audits, as well as end-users who have a right to know how their data is being used. Unfortunately, there are no established standards for this information, neither in terms of a canonical location nor format and content: Some websites provide traditional text documents via a clearly identifiable privacy policy link from their landing page, while others include this information in their Terms of Service. Some mobile applications employ buttons and icons to convey privacy information, while still others condense this information into more user-friendly privacy labels. Each of these user-facing formats has strengths while also presenting challenges for satisfying stakeholders' interests.

There are currently no commonly used consensus standards against which firms write their privacy documents. This lack of standardization leads to variability in their location, scope, format, and other features, which in turn creates difficulties for users, legislators, and other stakeholders to find, analyze, and act upon these documents. Of particular relevance is how to both manually and automatically find and evaluate documents for compliance.

4.1.1 Challenges

There are technical, social, and legal challenges to creating consensus around a standard for privacy documents, especially those that are end-user facing. Prior attempts at standardization have largely been seen as failures. We enumerate the challenges currently facing stakeholders and contextualize them in relation to past attempts at standardization.

4.1.1.1 Technical challenges: underspecified or volatile document features

There are a number of technical challenges to standardization. We enumerate the most salient of these below.

In this section we assume that a given product, system, or service requires a privacy document of some kind (i.e., we do not discuss the conditions under which such a product, system, or service would necessitate a privacy document). We abstract over *producers* of a document and *consumers* of a document; when the document feature entails different challenges for different elements of the producer-consumer relation, we ground these actors with specific examples.

Document location. There is no standard location where to find disclosures about a company's privacy practices, neither in terms of visual placement nor in terms of file path. For example, on the Web, links to a website's privacy policy are often placed in the website footer, placing a burden on the user to scroll down to find it. In mobile app stores, a link in the app listing should lead to the app's privacy policy, but often only leads to the developer's website, where the user has to conduct further investigation to find the policy. The problem is exacerbated if privacy information is made available in multiple languages. In these cases, the firm or service provider produces the privacy document, while the consumer may be an end-user, a bot or crawler, a lawyer, or even another producer (e.g., a company seeking to refer to a third party's privacy policy).

Dynamically generated web pages and symbolic paths complicate the automated localization of privacy documents. This is especially true for automated agents. That said, there are benefits to dynamic generation: rather than enumerating the full set of data practices, the producer can specialize their content to the particular product, service, jurisdiction, end-user, application, etc. This specialization benefits an end-user consumer, who only sees the relevant information. However, implementing this functionality correctly is quite challenging due to the range of components that the privacy document for an arbitrary producer-consumer pair may require.

Document granularity. The range of necessary information to include in a privacy document for a given producer-consumer pair points to the next challenge in developing standards: document granularity and scope. By "document granularity" we mean what fragment of a company's privacy disclosures are contained in a single document. By "scope" we mean the territorial, personal, and material boundaries the provisions of the privacy document are intended to apply to and the boundaries and extent of information that the policy covers regarding the collection, use, storage, and sharing of personal data. Scope can contribute to issues with granularity: some organizations have a single document named "Privacy Policy"; some incorporate this information into their Terms of Service, while others spread privacy information across multiple documents to make it more easily digestible or to adapt the presented information to different regulatory environments (special audiences, fields, or jurisdictions). Furthermore, organizations may provide separate privacy documents for each of their services. Finally, multiple pieces of software may have individual privacy policies that must be merged into a new service that uses them. There are currently no standard ways to compose these documents, leading to complex and illegible practices.

Document format. Privacy documents also widely differ in qualitative features that an end-user would experience. These include file type (e.g., PDF, HTML, plain text file), media type (text, tables, images, video), and subdivision into subsections and paragraphs. The information in privacy documents can also be visualized, summarized, and synthesized via non-

textual means of presentation, such as icons (e.g., the CCPA icon [1] or "nutrition labels" [2]). Privacy policy text also does not have to be static text but is sometimes dynamically generated. Beyond finding policy text, these differences in format and creation add additional obstacles to extracting privacy policy text for further processing and analysis [3].

Document content, audience, applicable jurisdiction, and scope. Format and presentation are closely related to the content of privacy documents, which in turn is influenced by layered regulatory requirements. Based on the location or jurisdiction alone, there can already be multiple tiers of regulation that apply, including the supranational (EU privacy laws), federal (national data protection laws), and state level (e.g., California Consumer Privacy Act (CCPA), Washington Privacy Act (WPA)). Other requirements hail from special regulations for protected audiences, such as the Children's Online Privacy Protection Act (COPPA), or specific fields or industries, such as the Health Insurance Portability and Accountability Act (HIPAA) for healthcare, the Gramm–Leach–Bliley Act (GLBA) for finance, or Family Educational Rights and Privacy Act (FERPA) for education.

While these regulations specify which content to include in privacy documents (e.g., the CCPA mandates that privacy policies need to state whether personal information is sold or shared for marketing purposes; the GDPR's disclosure requirements in Articles 13 and 14), they usually lack specific guidelines on how such content has to be presented.

4.1.1.2 Challenges of the standardization process

The technical challenges related to document location, granularity and scope, and format all impact any process of standardization. The need for standardization can be felt across different formats of privacy communication to facilitate adoption and address inherent differences in the objective of the communication. They should cover different modalities of communication, as well as the parties on either end of the communication. While the establishment of (privacy communication) standards at multiple points of a product's lifecycle may seem to be the key to richer communication, the fatigue of enforcing such standards could demotivate meaningful implementation. The challenge therefore lies in assessing the effective benefit of standardizing one or more points of communication, and whether the benefits will outweigh the complexity of the induced processes to meet such standards. This is amplified by the potential risk of introducing inconsistencies between different representations of a privacy concept across these multiple points. For example, this is applicable if different document variants are standardized for different stakeholders or when alternative visualizations are used to familiarize consumers about privacy practices. Vocabulary mismatches across standards, as well as in relation to regulations, can become failure causes in standards adoption (e.g., in P3P [7]).

The purpose behind data collection is ever evolving, can be generic in nature, and can be subject to a variety of legal and ethical frameworks. Under such situations, the requirements that a standard should meet are unclear. Standards can and do undergo revisions to incorporate the changing landscape of applicable platforms; however, the requirement to meet frequently revised standards may be seen as an operational burden and hamper adoption.

Furthermore, although it is understood that different stakeholders have different expectations from a privacy document, the precise nature of those expectations is understudied. Hence, what communication a standard should facilitate, and what structure is ideal for such communication, is likely to present itself as a challenge in standards development. Taxonomies help organize content in meaningful, unambiguous, and contained ways. The FPC states 11 Fair Information Practice Principles (FIPPs) for efficient privacy management [4]: consent

and choice; purpose legitimacy and specification; collection limitation; data minimization; use/retention/disclosure limitation; accuracy and quality; openness, transparency and notice; individual participation and access; accountability; information security; and privacy compliance. On the other hand, studies around privacy documents have focused on select categories of information: first party collection/use; third party sharing/collection; user choice/control; user access, edit & deletion; data retention; data security; policy change; Do Not Track; and international & specific audiences. These categories are loosely tied to the FIPPs and have served as a gold standard in data annotation and automated analysis attempts [6]. Nonetheless, the completeness of these categories in capturing all pertinent aspects of privacy communication, especially with respect to different stakeholders, is unknown. The lack of a well-crafted taxonomy creates barriers to standardization, which inherently must find ways to balance expressiveness and verbosity. On a similar note, a standardized vocabulary of privacy-relevant terms is also missing, which may lead to conflicting interpretations across specifications.

4.1.1.3 Sociopolitical challenges

A process of standardization can also be vulnerable to sociopolitical challenges. As learned from the P3P standardization process, this can manifest as waste of time while oscillating between specificity and generality, the introduction of too much transparency as viewed by specific stakeholders, the potential for misuse to generate a false notion of privacy, the introduction of the notion that a specification would substitute legislation, disparate stakeholders hinging on others to take the first step (who goes first – policy writers or policy checkers), and bare minimum implementations of a specification.

4.1.1.4 Technical challenges for firms attempting compliance

As consensus documents produced by formal organizations, privacy standards are typically designed to address legal obligations. These standards must then be translated into technical solutions. Additional challenges arise during this process. For instance, programmers face challenges when interpreting the content of a privacy policy against what a product or service actually does or against what it could do, especially when the product or service predates the policy.

4.1.2 Key research questions

What goals should a standardized privacy communication attempt to reach? Care has to be taken to account for the different expectations from involved groups (regulators, consumers, business owners), and accordingly preserve the indispensable elements. As such, identified objectives will significantly drive the design process of a standardized format, the specificity of the content used to realize a standard, and the feasibility of assessing if the objectives are met. The introduction of stakeholder-specific standards will inevitably introduce subsequent questions on standards mapping and also inform the creation of a comprehensive vocabulary and taxonomy of privacy communication artifacts. Some key research questions to consider include:

- What are the points in the data processing pipeline that could benefit from standards or format templates, and what is the smallest set of privacy data requirements to enable the functionality of a given service beyond that point?

- Where are there opportunities for auto-generating privacy documents, what kinds of formats should an auto-generator produce, and what entities and processes ought to govern the approval of new output formats?
- What inconsistencies can arise when multiple formats for the same concept are created, and how to address interpretive variations in these formats?
- What level of complexity of formal specification is required to capture the minimal expressiveness of different privacy document specifications?
- What quantitative and qualitative methods are needed to measure the conformance of an implementation to a specific standard, and what avenues exist (or need to be created) to integrate such conformance testing into an organization's operational activities? While quantitative methods are aimed at generating measurable insights into an implementation's adherence to concrete requirements of a specification (e.g., a data retention practice must unambiguously indicate a retention period; a purpose definition must be tied to every collected data artifact), qualitative methods are aimed at assessing conformance in terms of coverage, clarity, etc. Integration also includes the notion of feedback to facilitate iterative refinements.
- How to avoid sociopolitical and technical pitfalls in a standards development process? Recommendations such as simplicity and management of expectations are present from the P3P experience, but a deeper discussion is desired to prepare for previously unseen barriers.

4.1.3 Research directions

The specific solutions that resolve or address the challenges we enumerate should be in service of the following concrete desired outcomes:

- Identification of standardization points that serve (preferably) separate sections of an end-to-end privacy communication pipeline
- A systematization of objectives to be met at various standardization points
- A clear statement of requirements that implementation of the standard must meet: necessary, unambiguous, complete, precise, well-structured, consistent, testable [5]
- Formal methods to check for consistency violations in the standards; note: not only across revisions of a specific standard but also across manifestation of the same principle across standardization points
- A baseline implementation with guidelines for extensibility

Next, we discuss specific suggestions for the standardization of certain privacy document features.

Document location. This can be standardized by means of a known semantic endpoint for privacy-related information. Existing similar standards and proposals include:

- `robots.txt` (`https://www.rfc-editor.org/rfc/rfc9309`) – directives for bots which parts of a website (not) to access.
- `security.txt` (`https://www.rfc-editor.org/rfc/rfc9116.html`) – point of contact for security vulnerability notifications, already used by major companies.
- Other proposed standards of metadata files at well-known locations on web servers include `ads.txt`, `human.txt`, and `sellers.json`.
- There is already a proposal for a `privacy.txt` standard (https://privacytxt.dev), but unlike security or ads, the disclosures required in privacy policies are not universal and may vary between jurisdictions, which this proposal does not account for.

- Some companies already seem to acknowledge the problem of privacy documents being hard to find and automatically process, prompting them to provide a version of their privacy policy as a simple plaintext file; one example is Hulu (`https://www.hulu.com/privacy.txt`).

Document content and format. Templates can serve to capture high-level structures (e.g., prescribed headings, paragraphs, placements, etc.) as well as low-level structures (e.g., data items, data collector, purpose, trigger, mechanism, etc.).

Cross-referencing between multiple privacy documents. Privacy documents are hosted in a variety of formats, including privacy policy text, links, nutrition labels, and icons. Existing privacy documents often refer to each other. For example, a section in the general privacy document may refer to another, jurisdiction-specific document (e.g., designated privacy policies for residents of California, who are subject to the CCPA). These cross-references complicate human understanding of the issuer's exact data practices and their associated privacy rights. Designing a standardized path through the documents, links, labels, and icons would create a hierarchical organization of discrete, distributed documents that would aid different stakeholders in finding the applicable privacy information.

References

1 California Department of Justice. *CCPA Privacy Icons*. `https://oag.ca.gov/privacy/ccpa/icons-download`, accessed May 21, 2025.

2 Apple Inc. *App Privacy Details*. `https://developer.apple.com/app-store/app-privacy-details/`, accessed May 21, 2025.

3 H. Hosseini, M. Degeling, C. Utz, and T. Hupperich. *Unifying Privacy Policy Detection*. Proceedings on Privacy Enhancing Technologies (PoPETs), 2021(4), pp. 480–499. `https://doi.org/10.2478/popets-2021-0081`

4 ISO/IEC. *Information technology-Security techniques-Privacy framework*. International Organization for Standardization, Geneva, Switzerland, International Standard ISO/IEC 29100:2011(E), 2011.

5 ETSI. *A Guide to Writing World Class Standards*. `https://www.etsi.org/images/files/Brochures/AGuideToWritingWorldClassStandards.pdf`, accessed May 21, 2025.

6 S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell et al. *The creation and analysis of a website privacy policy corpus*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1330–1340.

7 A. Schwartz. *Why P3P didn't work*. `https://cdt.org/wp-content/uploads/pdfs/P3P_Retro_Final_0.pdf`, accessed May 21, 2025.

4.2 Datasets, Automation, and Analysis Methods

Peter Story (Clark University – Worcester, US), Sepideh Ghanavati (University of Maine, US), Henry Hosseini (Universität Münster, DE & Westfälische Hochschule – Gelsenkirchen, DE), Jelena Mitrovic (Universität Passau, DE & Institute for Artificial Intelligence R&D of Serbia – Novi Sad, RS), Tim Samples (University of Georgia, US), Isabel Wagner (Universität Basel, CH), and Tianyang Zhao (Pennsylvania State University – University Park, US)

4.2.1 Challenges

Concerning datasets, automation, and analysis methods for privacy documents, we have identified four key challenges as well as four meta-challenges. These challenges deliberately focus on the landscape of privacy documents found today, instead of considering possible future improvements or standardization efforts. The rationale behind this focus is that the current state is likely to persist for at least several years, given the typical duration of legislative processes (e.g., the GDPR was first proposed in 2012 and came into effect in 2018) and the current lack of substantial legislative initiatives regarding privacy documents in the EU and elsewhere.

The primary stakeholders in this research area are fellow privacy researchers, as well as regulatory authorities, end users, and developers.

The first challenge is *continuously collecting privacy documents from many companies in many languages*. This is important for providing training data for classifiers, enabling the tracking of changes in policy documents, and ensuring greater inclusivity with respect to languages and jurisdictions.

The second challenge is about *scaling the automated annotation of privacy documents*. For example, regarding the number of annotations, effort, or standard labeling schemes, this approach aims to reduce manual effort and enable more large-scale annotations to train more capable classifiers.

The third challenge is developing *effective computational methods* to analyze privacy documents. This enables the building of tools for stakeholders, including summarization, locating opt-outs, detecting inconsistencies, assigning privacy grades, and identifying topical trends over time.

The fourth challenge is *detecting inconsistencies between software, privacy labels, and privacy and policy documents*. This is important to ensure compliance with laws and avoid misleading users about the privacy practices of systems.

In addition to the four challenges described above, we further identified four meta-challenges, which we describe next, that broadly apply to some or all of the challenges we identified in the area of datasets, automation, and analysis methods. Failing to address these meta-challenges risks limiting the longevity, applicability, impact, and reproducibility of research.

The *FAIR principles* [1] (findability, accessibility, interoperability, and reusability) apply primarily to collected datasets, such as corpora and annotations, but the spirit of the principles can also inform the development of solutions for privacy policy analysis, such as models and tools. Considering the FAIR principles can, for example, influence the choice of file formats. Plain-text formats, such as CSV, should be preferred over proprietary file formats, such as Excel, or formats that require infrastructure, such as an SQL server. The FAIR principles can also inform the instructions given to annotators (e.g., providing codebooks) and the detail and structure of documentation provided with datasets (e.g., including each coder's annotations in addition to the agreed-upon annotations).

Environmental impact and sustainability are rapidly becoming important considerations in science, especially in light of the resource consumption of new machine learning approaches such as large language models (LLMs) and the potential impact on climate change. Large-scale research on privacy documents could lead to substantial resource consumption. Resource consumption should be measured and reported from the start. When the deployment phase of systems is reached, the insights derived from this information can assist in evaluating the balance between task efficiency and the utilization of resources, guiding decisions on potential tradeoffs.

The *selection of downstream tasks* for analyzing privacy documents should be prioritized. Privacy documents contain a wide variety of information, and research has proposed methods to extract and analyze different aspects of this information. These downstream tasks range from broad classification of information categories in privacy policies to identifying narrow information items, such as opt-out statements, to analyzing the consistency of policy statements with data flows. While working on novel downstream tasks may be beneficial for academic indicators of success, it is also crucial to consider which downstream tasks are most beneficial for stakeholders, such as end users.

Although the top privacy publication venues, such as IEEE S&P and USENIX Security, are often novelty-driven, *maintainability and availability of research products* should be given more attention. Past research projects have generated numerous artifacts, such as tools and corpora, that could be useful and applicable to other research projects. However, research projects are often forced to duplicate or repeat previous efforts because tools and datasets are not maintained, unavailable, or insufficiently documented. The challenge is that maintaining research products is not well incentivized in the academic world and is therefore not a natural outcome of academic work compared to industry products. In the interest of open science and reproducibility, research projects should plan for long-term maintainability and availability of their research products from the start, e.g. by selecting repositories and platforms that ensure long-term availability (e.g., preferring Zenodo and Software Heritage over self-hosted websites) or by committing technician and/or student time to maintenance.

4.2.2 Key research challenges and questions

4.2.2.1 Challenge 1: Continuous multilingual large-scale collection, storage, and organization of privacy documents

- RQ1.1: How can we improve the automated collection, storage, and organization of privacy documents?
- RQ1.2: Which privacy document(s) apply to a system? Some documents reference other documents, sometimes from other companies. For example, a mobile app's privacy policy may reference the privacy policy of a third-party analytics library.
- RQ1.3: Which sections of a privacy document apply to a system? For example, a company that offers a wide range of products may write a single privacy policy that covers all of its products. Additionally, certain sections of a privacy document may only apply to children or vulnerable groups, or may only be applicable in specific jurisdictions.
- RQ1.4: How prevalent are non-textual elements in privacy documents? Privacy documents are often reduced to plain text for analysis, while this approach may result in the disordering of textual fragments and the loss of information for certain elements. Examples of such potentially problematic elements include tables, multi-layered policies, lists, etc.
- RQ1.5: Can we generate changelogs to highlight privacy document changes? Some companies visualize document changes using a "diff" (e.g., Google). It might be possible to generate diffs for the policies of other companies. In addition, it may be useful to go beyond a simple diff, perhaps highlighting or summarizing "interesting" changes.

4.2.2.2 Challenge 2: Scaling annotation of privacy documents

- RQ2.1: Can we use large language models (LLMs) to help scale the annotation of privacy policies, with quality comparable to human annotators?
- RQ2.2: How can we improve annotation practices for privacy documents? What are the pros and cons of various annotation tools? How can annotation tools be improved?
- RQ2.3: How can we create annotated corpora in multiple languages, which also meet the regulatory standards in various jurisdictions?

4.2.2.3 Challenge 3: Developing effective computational methods to analyze privacy documents

- RQ3.1: What computational methods offer the greatest performance for downstream tasks? Downstream tasks may involve extracting information from privacy documents, such as opt-out choice links, types of information collected and shared, data retention duration, and data deletion options.
- RQ3.2: Which computational methods balance task performance with other goals? An example of such balance could be reducing resource consumption while improving explainability.
- RQ3.3: What privacy-relevant content is present in documents other than designated privacy policies? For example, terms of service, cookie policies, cookie banners, community guidelines, and FAQs may contain information that is relevant to privacy.

4.2.2.4 Challenge 4: Detecting inconsistencies between software, privacy labels, and privacy documents

- RQ4.1: How can we determine privacy-related behaviors of software systems? These could include, among others, mobile applications, web applications, IoT devices, smart cities, or vehicles.
- RQ4.2: What methodologies can be employed to identify privacy-related behaviors within a given source code?
- RQ4.3: In what ways can we assist developers with limited resources in generating privacy documents, while ensuring alignment and consistency between the written text and the corresponding code?

4.2.3 Research directions

4.2.3.1 Addressing challenge 1

To address RQ1.1, while the toolchain by Hosseini et al. [2] provides a comprehensive solution to collect and preprocess privacy documents automatically on a large scale in 41 languages, the final stage of this toolchain, which uses trained classifiers to distinguish between privacy documents and non-privacy documents, is limited to English and German. Additional classifiers that cover the 39 other languages that the toolchain can collect and preprocess would improve inclusivity.

To address RQ1.2, open web indices, such as the OpenWebSearch.eu Open Web Index (OWI)[1], can be used to identify linking relationships to and from privacy documents [9]. As part of the preprocessing of a privacy document, any referenced privacy documents can be retrieved and inserted into the "main text" of the original privacy document.

[1] https://openwebsearch.eu/open-webindex/

To address RQ1.3, classifiers can be developed to identify privacy-relevant sections in related documents (e.g., terms of use). These sections could be included in the overall privacy analysis of the system.

To address RQ1.4, a measurement study could be conducted to determine the prevalence of problematic elements that hinder machine-readability of privacy documents (e.g., tables, multi-layered policies, enumeration). Data from this study could inform the development of tools for converting problematic elements to easier-to-parse plain text for classifiers.

For RQ1.5, while the GDPR mandates that users be informed about changes in privacy policies that alter the legal basis or purpose of processing, not all companies may comply with this requirement. A simple "diff" can be generated from the plain text versions of privacy policies. However, to make these results useful to users, it is necessary to develop automated methods to filter out insignificant changes (e.g., rephrasing) and to highlight the impactful changes.

4.2.3.2 Addressing challenge 2

To address RQ2.1, approaches that utilize LLMs in an active learning setting have been demonstrated to be effective in numerous NLP tasks [11]. LLM-in-the-loop approaches should be attempted for privacy documents if the format and annotation schema allow. When exploring approaches using LLMs, it is essential to ensure the reproducibility of the results. Sharing source code, prompts, and the checkpoint/version of LLMs is essential, as some LLMs are not open-source and do not always show the latest version (e.g., ChatGPT). It should be standard practice to report the resources used, including running time and computational resources such as GPUs.

Researchers should conduct a comprehensive study that compares the quality of such machine-assisted annotations with that of human-annotated corpora. The results may depend on the type of annotation schema used, so it would be important to study multiple annotation schemes.

To address RQ2.2, researchers can start by creating and maintaining an online resource (e.g., a wiki) describing annotation tools and their pros and cons. It may also be worth improving existing tools, such as using simple NLP methods to highlight negations in document text. An annotation SoK paper could also be published to highlight annotation best practices. For example, it is essential to develop a usable annotation scheme that other labs can apply with low error rates. One approach is to limit the annotation scheme to a single page to limit its complexity. A challenging aspect of annotation is handling disagreements between annotators [3]. Current customs include majority vote and union of labels. With human annotators, we can take the approach of hosting a meeting at the end to resolve disagreements.

Apart from extending the size of the current corpora, the need for regulatory-aware collections of annotated documents arises, especially in Europe, where the GDPR is applicable. For example, in a corpus annotated using a GDPR-based annotation schema [10], what are the relations between specific data processors and data controllers *across the entire corpus*? A possible approach to addressing this question could be the creation of a graph-based collection of documents that contains relations between paragraphs of privacy documents, as well as information on how different labels relate to one another.

RQ2.3 can be explored through large-scale targeted crawling campaigns. These could extend existing web-based corpora (e.g., the English-German corpus by Arora et al. [4]) and enhance already developed crawlers [2], as well as utilize already existing web indices to filter out more task-specific documents and create large, multilingual corpora of privacy documents

that would then still require labeling. Another approach could be to leverage open machine translation software. However, one needs to be aware of the different jurisdictions and create relevant labels for each language. Finally, it is essential to develop taxonomies that are applicable to various research questions and regulatory environments. This would ensure that datasets have lasting value in the face of evolving regulations.

4.2.3.3 Addressing challenge 3

To address RQ3.1, researchers should compare the relative task performance of different computational methods for analyzing privacy documents. For example, comparing the accuracy of text classification using LLMs with classical machine learning algorithms such as logistic regression (LR). It is also worth exploring hybrid techniques, such as symbolic NLP in combination with LLMs. To report task performance, researchers should report at a minimum the accuracy score, the F1 score, and the number of ground truth instances in each category. Other metrics, such as precision and recall, are also beneficial. Researchers can also perform an ablation study to determine which features are the most important.

To address RQ3.2, researchers should report task performance metrics (e.g., accuracy) alongside other quality metrics. Quality metrics may include computational resource consumption, licensing costs, and explainability. Depending on the downstream task, task performance might be more or less important than other quality metrics. For example, if LLMs and LR offer similar task performance, then LR might be preferable due to its lower resource consumption and greater explainability. The predictions of an LR model can be understood by examining the model coefficients. In contrast, if LLMs perform tasks substantially better than LR, LLMs might be chosen despite consuming more computational resources, having higher licensing costs, and offering limited explainability. Of course, the choice between models is context-specific and would depend on the downstream task. Another factor to consider is that the use of closed-weight LLMs (e.g., ChatGPT, Gemini, Claude) may limit the reproducibility of research. Closed-weight LLMs can be modified server-side without notice, or access could be completely revoked. In contrast, open-weight LLMs (e.g., Llama) can be archived, which supports reproducible research. We recommend that researchers perform tasks using open-weight LLMs, perhaps in conjunction with closed-weight LLMs. Access to model weights will also affect the deployment for downstream tasks.

To address RQ3.3, researchers should develop computational methods for documents other than privacy policies. Terms of service, cookie policies, cookie banners, community guidelines, and FAQs may all contain privacy-relevant information. Privacy policies often fail to address users' privacy-related questions. However, answers to users' questions might be found in other documents, such as privacy FAQs. Thus, tools to answer users' questions will be more effective if they can draw from a variety of privacy-related documents.

4.2.3.4 Addressing challenge 4

To address RQ4.1, several directions could be followed. For example, researchers could use crowdsourcing to collect data from users about applications, thereby inferring backend data-handling practices. In addition, reverse engineering techniques could be used to understand the behavior of closed-source applications [5]. Lastly, researchers could advocate for regulatory bodies to provide them with access to source code, thus enhancing the privacy and security of the general public. This is similar to how the EU AI Act envisions access to models, algorithms, and datasets.

To address RQ4.2, researchers should examine the source code of various applications to create a comprehensive taxonomy of privacy behaviors. Past research [6, 7, 8], for example, defined privacy behaviors in terms of four categories of practices (i.e., collection, sharing, processing, and others) and four categories of purposes (i.e., functionality, advertisement, analytics, and others). These categories are limited and do not encompass various cases of privacy behaviors beyond those related to permissions and access. Researchers could begin with the existing taxonomies for policy documents and extend them to source code. Using the expanded taxonomy, researchers should focus on creating ground truth datasets. Creating such datasets can be cumbersome, but with the advancement of LLMs, these models can be used in conjunction with human-in-the-loop (HitL) approaches to create more robust ground-truth datasets. These datasets also require evaluation. Potentially, independent developers could be recruited through crowdsourcing platforms to assess and improve the quality of the dataset.

To address RQ4.3, researchers should leverage, extend, and develop new static or dynamic analysis tools to identify and extract data flows from the source code. It is worth going beyond mobile applications and considering other software, such as backend code. Researchers must determine the most effective way to prepare and represent code for model training and inference. Approaches that focus on pruning and slicing source code to focus on privacy features should be explored. To identify inconsistencies between source code and privacy policies, researchers should focus on mapping and creating traceability between source code and policy documents or labels. Lastly, software engineering research has focused on code summarization and captioning for more than a decade. Researchers in the privacy domain could leverage or adopt some of the techniques from software engineering to automatically generate labels from policy texts and code, and do the translation.

References

1 M. D. Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.

2 H. Hosseini, C. Utz, M. Degeling, and T. Hupperich. *A Bilingual Longitudinal Analysis of Privacy Policies Measuring the Impacts of the GDPR and the CCPA/CPRA*. Proceedings on Privacy Enhancing Technologies, 2024(2):434—463, February 2024.

3 D. G. Gordon and T. D. Breaux, *The role of legal expertise in interpretation of legal requirements and definitions*. 2014 IEEE 22nd International Requirements Engineering Conference (RE), Karlskrona, Sweden, 2014, pp. 273-282, doi: 10.1109/RE.2014.6912269.

4 S. Arora, H. Hosseini, C. Utz, V. B. Kumar, T. Dhellemmes, A. Ravichander, P. Story, J. Mangat, R. Chen, M. Degeling, T. Norton, T. Hupperich, S. Wilson, and N. Sadeh. *A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus*. In Proceedings of the 13th Conference on Language Resources and Evaluation, LREC 2022, pages 5460–5472, Paris, France, 2022. ELRA.

5 S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh. *MAPS: Scaling Privacy Compliance Analysis to a Million Apps*. Proceedings on Privacy Enhancing Technologies, vol. 2019, no. 3, pp. 66–86, Jul. 2019, doi: 10.2478/popets-2019-0037.

6 V. Jain, S. Ghanavati, S. T. Peddinti, and C. McMillan. *Towards Fine-Grained Localization of Privacy Behaviors*. In Proceedings of the 8th IEEE European Symposium on Security and Privacy (Euro S&P'23), Delft, July 3-7, 2023.

7 V. Jain, S. D. Gupta, S. Ghanavati, S. T. Peddinti, and C. McMillan. *PAcT: Detecting and Classifying Privacy Behavior of Android Applications*. In Proceedings of the 15th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '22), ACM, NY, USA, 104–118, 2022.

8 V. Jain, S.D. Gupta, S. Ghanavati, and S.T. Peddinti. *PriGen: Towards Automated Translation of Android Applications' Code to Privacy Captions*. 15th International Conference on Research Challenges in Information Science (RCIS2021), Cypress, 2021.

9 M. Granitzer, S. Voigt, N.A. Fathima, M. Golasowski, C. Guetl, T. Hecking, G. Hendriksen, D. Hiemstra, J. Martinovič, J. Mitrović, I. Mlakar, S. Moiras, A. Nussbaumer, P. öster, M. Potthast, M. Srdič Senčar, M. Sharikadze K. Slaninová, B. Stein, A. de Vries, V. Vondrák, A. Wagner, and S. Zerhoudi. *Impact and development of an Open Web Index for open web search*. Journal of the Association for Information Science and Technology, vol. 75, no.5, pp. 512-520, 2024, doi:10.1002/asi.24818.

10 H. Darji, J. Mitrović, and M. Granitzer. *German BERT Model for Legal Named Entity Recognition*. Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART) vol. 3, pp. 723-728, 2023. doi:10.5220/0011749400003393.

11 N. Kholodna, S. Julka, M. Khodadadi, M.N. Gumus, and M. Granitzer. *LLMs in the Loop: Leveraging Large Language Model Annotations for Active Learning in Low-Resource Languages*. In Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track. ECML PKDD 2024. Lecture Notes in Computer Science, vol. 14950. pp. 397-412. doi:10.1007/978-3-031-70381-2_25.

4.3 Usable and Useful Notice & Consent

Simone Fischer-Hübner (Karlstad University, Chalmers University of Technology and Gothenburg University, SE), Kai-Wei Chang (UCLA, US), Nico Ebert (ZHAW – Winterthur, CH), Agnieszka Kitkowska (Jönköping University, SE), and Shidong Pan (Australian National University – Canberra, AU)

4.3.1 Challenges

4.3.1.1 Challenge 1: Limitations of human-centered and inclusive approaches to the design of notice and consent

Better stakeholder engagement. In the design of notice and consent, not only end-users but various other stakeholder groups that engage with privacy practices and consent mechanisms need to be actively included and involved in the development and design process. These can be lawyers, software engineers, and other stakeholder groups, such as journalists or any marginalized or vulnerable groups.

Better integration of diverse communication channels, formats, and media. A holistic and inclusive approach is needed that also considers a variety of communications and interaction channels and formats. Due to emerging technology (e.g., AI-based conversational agents, voice assistants, autonomous vehicles, and VRs that intrusively collect large amounts of more sensitive personal data), new ways for effective, usable and useful notice and consent are required (e.g., audio-assisted consent). With an increasing emphasis on inclusivity also additional design aspects are important, for instance, complementary easy-accessible audio formats of privacy information should be available for people with visual impairments. Similarly, the intersectional lens could be considered, where, in the given example, also easy to comprehend language is used in the audio-notice to ensure the inclusion of people with lower cognitive functions or migration backgrounds. For example, a privacy notice in audio

format might use straightforward terms such as "we use your email to send updates about our products" instead of complex legal jargon like "we process your contact information to disseminate product-related notifications".

Better personalization and contextualization of notice and consent (including language). There is a growing need for better personalization in notice and consent to reflect the individual demands and preferences of stakeholders. People from diverse backgrounds may rely on different terminologies or expressions. For instance, while a lawyer might understand a term like "third-party", it would likely be incomprehensible to a regular user. Adapting language and presentation to suit the broader audience is crucial for improving comprehension. Moreover, individuals typically may have different preferences regarding policy content that are of interest or relevance to them. To ensure improved personalization, there is a need for a stronger focus on user context to make privacy notice and consent more relevant and engaging for the different groups of users. While theories (e.g., contextual integrity) and new approaches that respect contexts have been developed (e.g., contextual privacy notices [16], just-in-time notices [19]), further research is needed, in particular considering the contexts within the emerging technologies (e.g., VR, Metaverse, or smart technologies).

In summary, there is a need for a human-centered and inclusive design approach that considers a variety of stakeholders and their preferences, channels/formats and personalization.

4.3.1.2 Challenge 2: A lack of a risk-oriented approach to the design of notice and consent

What is presented in a notice and consent process typically refers, in general terms, to all possible data processing practices, data controllers, data processors, and other third parties ("procedural transparency"). Users might be confronted with too much, too irrelevant, or deceiving information. At the same time, potential risks of data processing might not be transparent and clear to them. As dealing with privacy information is often only a secondary task for users, a focus on privacy risks could help to make communication more effective ("risk transparency"). A risk-based approach to communication has been proven effective in other areas, such as safety, and indications for their effectiveness in privacy exist, too [1]. It can be seen as complementary to traditional approaches that want to give a complete picture of data processing practices and will remain relevant for specific stakeholder groups (e.g., privacy-interested users or professionals).

A risk-oriented approach would, however, require identifying what privacy risks exist and how they can be categorized. It would also require new tools (e.g., based on AI) that help to quickly predict risks and allow dynamic risk communications (e.g., based on changes in the privacy policy that introduce new or unexpected risks and in the context of dynamic consent).

Implementing a risk-oriented approach in privacy notices could foster end-user trust in the respective organization, if it is also clearly communicated if and how risks are adequately mitigated. A challenge is communicating residual risks in relation to the benefits of disclosing data, e.g., when using a service. This could not only provide increased transparency for users but also for controllers, who could see benefits in a transparent risk-based approach that can also create trust and, therefore, would support it.

4.3.1.3 Challenge 3: A lack of standardized methods to evaluate the usability and usefulness of notice and consent

Although researchers typically evaluate new design artifacts, it is difficult to compare the results of different studies with regard to general criteria. Standardized design criteria and evaluation methods could help to assess the benefits of artifacts. A concrete example is a validated user-survey instrument comparable to the system usability scale [11] that would allow the community to compare different artifacts (e.g., nutrition labels vs. short privacy notices with complementing privacy icons). There are also no standardized methods for the evaluation across contexts or channels/formats (e.g., privacy information presented in a car vs. on a mobile phone). Standardized methods could range from general criteria for "good" design of notice and consent to validated instruments to survey users' perceptions of the usability and usefulness of notice and consent.

Another issue exacerbating the problem of the lack of standardized evaluation methods in the space of design of notice and consent is lack of reproducibility (when the measurement can be obtained with stated precision by a different team, a different measuring system, or in a different location on multiple trials), replicability (when the measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials) and repeatability (when the measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials) of research. Not only are such studies scarce within the field, but also difficulties related to how such research is defined should be addressed, considering contextual factors surrounding privacy (e.g., temporality). For example, while some researchers may claim the effectiveness of newly proposed notice and consent artifacts, other researchers may have no incentives, capability, or means to reproduce the results.

Also, structural issues (e.g., lack of outlets accepting replication studies) may prevent replication. Furthermore, the introduction of AI components adds another layer of complexity, as the inherent uncertainty of AI-driven systems can make reproducibility even more challenging. Replication, reproduction, and repetition, however, are necessary to create an actual body of knowledge in the discipline. Adding to this challenge are the often little interest and encouragement from publication venues towards publishing replication studies, as well as requiring researchers to make the datasets, artifacts, and other research-relevant materials publicly available.

4.3.2 Key research questions

The overarching goal to achieve in the next decade is to define guidelines enabling the holistic human-centered approach to the design of notice and consent. Specifically, we have to address the following:

- RQ1: How should a holistic human-centered approach to the design of consent and notice look like – and how can it address key design challenges?
- RQ2: How can privacy risks be identified and communicated in a way that benefits users and other stakeholders and fosters reliable trust?
- RQ3: How can we standardize methods for the evaluation of usability and usefulness of privacy policies across research?

Table 1 draws connections between the identified challenges, research questions, and suggested approaches.

4.3.3 Research Directions

In order to develop solutions we could adapt approaches from related areas to our problem domain.

4.3.3.1 Approach 1: Human-centered and inclusive privacy by design and default

Usability is an important prerequisite for privacy by design and by default, and at the same time privacy by design and by default provides means for enhancing usability. Moreover, solutions for human-centric privacy by design and by default have to be inclusive and adapt to the needs and values of different types of users and other relevant stakeholders. This approach helps to balance different design requirements in the design space for privacy notices [19]. Hence, new methodologies and approaches for privacy by design (for privacy notices and consent solutions) based on and combined with human-centric and inclusive design approaches should be researched and developed [8].

4.3.3.2 Approach 2: Risk perception and risk communication

Risk research has a long tradition of studying risk perception and risk communication in the non-digital world that could also be applied in traditionally digital domains [22, 21]. Also in digital domains, for example, "saliency" of privacy risks plays a key role in improving risk perception [6] and promoting protective user behavior [7]. Other ideas from risk research have not yet found their way into privacy research but might be beneficial in developing risk-oriented approaches in the domain of notice and consent. For example, risk researchers hypothesized that risk mitigation measures may lead to an increased risk acceptance level of individuals ("risk compensation" [9]). For instance, explanations of privacy-enhancing technologies (PETs) could introduce an unexpected level of personal data sharing on the side of the end-user when all risks are assumed to be mitigated. The opposite effect could occur if the core protection functionality of PETs is misunderstood. Further, different theoretical lenses could be applied to identification of privacy risks, particularly among marginalized populations. One approach could be critical feminist frameworks, such as intersectionality, that have been used in the field of human-computer interaction (HCI) [20].

4.3.3.3 Approach 3: AI support and tools

Artificial intelligence (e.g., LLMs) may provide means that can be beneficial in all phases of the design process, from creation to evaluation of artifacts, but also as a functional part of proposed design artifacts. In the design phase, an LLM can be used to simulate different personas with different privacy preferences or groups of people to quickly evaluate privacy design in an early stage. In the consent interpretation stage, recent years have seen the development of machine learning-based Personalized Privacy Assistants (PPAs) that can, based on an analysis of the users' previous privacy decisions (e.g., related to setting or rejecting privacy permissions for Android or IoT systems [12, 2, 23, 3]), predict the users' preferred choices and subsequently assist users in privacy decision-making with suitable recommendations. Nonetheless, PPAs can only help to semi-automate privacy decisions like consent, which according to the GDPR requires an affirmative action by the user and thus cannot be fully automated [15]. Still, in the future, PPAs could be developed more broadly for other technical areas (e.g., for IoT Trigger Action Platforms, cloud environments) and also assist users, e.g., with extracting or highlighting core information to meet their personal interests in addition to the personalized recommended decisions. In the consent

Table 1 Research Questions, Challenges, and Approaches.

Research Question	RQ/Challenge	Approach
What immediate research questions need to be answered?		
How can relevant stakeholders be identified and engaged in the phases of requirement elicitation and the generation, design, and evaluation of privacy notices and consent?	RQ1/C1	A1
What are more effective user interactions with policy content for raising the user's attention and awareness of the relevant policy information, esp. in the context of consent (e.g., via interactive voice communication)?	RQ1/C1	A1
What are relevant risks from a user perspective, and how can risk communication be best personalized with the help of AI tools?	RQ2/C2	A2
What tools do developers need to be better supported in the development of privacy notices and consent?	RQ3/C3	A4
What best practices can be proposed for replication studies on usable privacy related to privacy notice and consent?	RQ3/C3	A4
What are the challenges and questions to be answered in the next three years?		
How can we apply AI to facilitate the design and produce personalized, useful privacy notices for users?	RQ1/C1	A3
What are usable and inclusive forms of policy communication utilizing multiple channels?	RQ1/C1	A1
How can transparency about risks and perceived consequences be achieved without unnecessarily scaring users?	RQ2/C2	A2
How can risks, risk mitigation and the residual risks be made more transparent?	RQ2/C2	A2
How can personal or societal benefits achieved from disclosing personal data compared to the residual risk be communicated?	RQ2/C2	A2
How can risks be dynamically detected and used to inform the users and obtain dynamic consent?	RQ2/C2	A3
How effective is a risk-based approach to privacy communication compared to traditional approaches?	RQ2/C2	A3
What are challenges and questions to be answered within the next decade?		
What are the criteria for the evaluation of the usability and usefulness of privacy policy notice and consent?	RQ3/C3	A4
What "evaluation checklist" and tool support for privacy notice and consent should be provided to developers?	RQ3/C3	A4

informing stage, AI-based tools or PPAs could be used to create personalized and dynamic privacy information that adapts to users' context and information needs [12]. For instance, if a weather app starts to use location data not only for the purpose of showing the local weather but for location-based advertising or sharing data with location data brokers, LLMs can easily generate a new consent prompt to describe the data practice, and users could be notified and requested to consent dynamically. At the same time, such mechanisms need to be designed to be privacy- and values-preserving, legally compliant, and ethical [15, 18, 17]. This is particularly important given the widespread criticism of AI techniques for their frequent hallucinations (in the case of LLMs) and lack of transparency.

4.3.3.4 Approach 4: Systematic reviews, replication studies, and mega-studies

To enable the creation of standardized evaluation criteria that could be applied in the research on privacy notice and consent design, there is a need for the production of more empirical evidence within the field. However, to identify what empirical investigations are more urgent in the fast-changing technology landscape, more systematic reviews and meta-analysis studies are needed. Only based on such studies can researchers pursue the right research problems.

Moreover, considering the evaluation criteria and assessment of the effectiveness of design, the field could pursue methods used in the social sciences, such as mega-studies [4]. Such studies were shown to be successful in making behavioral science findings more successful in their applicability, particularly in the context of behavior change through design in the context of nudging [5, 10, 13, 14]. Mega-studies themselves, although challenging to conduct, could also serve as quantitative evaluation criteria.

In order to develop solutions, we would also require additional skills and collaboration with other fields. Research on a holistic human-centered approach to the design of privacy notice and consent will require interdisciplinary expertise and the cooperation of a broad range of stakeholders. Most importantly, the research needs to include experts from the fields of computer science, information systems, social science, economics, psychology, risk research, safety, and law and should also be based on requirements collected from stakeholders elicited from end users with different demographic backgrounds, representatives from organizations including management and DPOs, as well as regulators.

References

1 Almuhimedi, Hazim, et al. "Your location has been shared 5,398 times! A field study on mobile app privacy nudging." Proceedings of the 33rd annual ACM conference on human factors in computing systems. 2015.
2 Bahirat, Paritosh, et al. "A Data-Driven Approach to Developing IoT Privacy-Setting Interfaces", 23rd International Conference on Intelligent User Interfaces, pp. 165-176, March 2018, [online] Available: https://dl.acm.org/doi/10.1145/3172944.3172982.
3 Das, Anupam, et al. "Personalized privacy assistants for the internet of things: Providing users with notice and choice." IEEE Pervasive Computing 17.3 (2018): 35-46.
4 Duckworth, Angela L., and Katherine L. Milkman. "A guide to megastudies." PNAS nexus 1.5 (2022): pgac214.
5 Duckworth, Angela L., et al. "A national megastudy shows that email nudges to elementary school teachers boost student math achievement, particularly when personalized." Proceedings of the National Academy of Sciences 122.13 (2025): e2418616122.
6 Ebert, Nico, et al. "Bolder is better: Raising user awareness through salient and concise privacy notices." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021: 1–12.

7 Ebert, Nico, et al. "When information security depends on font size: how the saliency of warnings affects protection behavior." Journal of Risk Research 26.3 (2022): 233–255.

8 Fischer-Hübner, Simone, and Karegar, Farzaneh. "Addressing Challenges: A Way Forward." The Curious Case of Usable Privacy: Challenges, Solutions, and Prospects. Cham: Springer International Publishing, 2024. 133-160.

9 Hedlund, James. "Risky business: safety regulations, risk compensation, and individual behavior." Injury prevention 6.2 (2000): 82-89.

10 Kuan, Robert, et al. "Behavioral nudges prevent loan delinquencies at scale: A 13-million-person field experiment." Proceedings of the National Academy of Sciences 122.4 (2025): e2416708122.

11 Lewis, James R. "The system usability scale: past, present, and future." International Journal of Human–Computer Interaction 34.7 (2018): 577-590.

12 Liu, Bin, et al. "Follow my recommendations: A personalized privacy assistant for mobile app permissions." Twelfth Symposium on Usable Privacy and Security (SOUPS 2016). 2016.

13 Milkman, Katherine L., et al. "Megastudies improve the impact of applied behavioural science." Nature 600.7889 (2021): 478-483.

14 Milkman, Katherine L., et al. "A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment." Proceedings of the National Academy of Sciences 118.20 (2021): e2101165118.

15 Morel, Victor, and Fischer-Hübner, Simone. "Automating privacy decisions-where to draw the line?" 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). IEEE, 2023.

16 Pan, Shidong, et al. "A NEW HOPE: Contextual Privacy Policies for Mobile Applications and An Approach Toward Automated Generation" 33rd USENIX Security Symposium (USENIX Security 24). 2024.

17 Pan, Shidong, et al. "Is It a Trap? A Large-scale Empirical Study And Comprehensive Assessment of Online Automated Privacy Policy Generators for Mobile Apps." 33rd USENIX Security Symposium (USENIX Security 24). 2024.

18 Morel, Victor, et al. "AI-driven Personalized Privacy Assistants: a Systematic Literature Review." (2025), `https://arxiv.org/abs/2502.07693`

19 Schaub, Florian, et al. "A design space for effective privacy notices." Eleventh Symposium on Usable Privacy and Security (SOUPS 2015).

20 Schlesinger, Ari, et al. "Intersectional HCI: Engaging identity through gender, race, and class." In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 5412-5427).

21 Siegrist, Michael, & Árvai, Joseph (2020). Risk perception: Reflections on 40 years of research. Risk Analysis 40(S1) (2020), pp. 2191-2206.

22 Slovic, Paul, et al. Why study risk perception?. Risk Analysis, 2(2) (1982), pp. 83-93.

23 Smullen, Daniel, et al. "The Best of Both Worlds: Mitigating Tradeoffs Between Accuracy and User Burden in Capturing Mobile App Privacy", Proceedings on Privacy Enhancing Technologies, 2020(1), pp. 195-215.

4.4 Consumer Privacy Beyond Notice and Choice

Noah Apthorpe (Colgate University – Hamilton, US), Eleanor Birrell (Pomona College – Claremont, US), Travis Breaux (Carnegie Mellon University – Pittsburgh, US), Kirsten Martin (Carnegie Mellon University – Pittsburgh, US), Rishab Nithyanand (University of Iowa – Iowa City, US), Sarah Radway (Harvard University – Allston, US), Yan Shvartzshnaider (York University – Toronto, CA), and Maximiliane Windl (LMU München, DE)

4.4.1 Challenges

The notice and choice regime that appears in US and EU privacy law has dominated how online privacy is regulated. Websites, apps, IoT devices, and other technologies post privacy policies describing (to some degree) data practices, and people are expected to choose services that meet their privacy needs.

However, decades of research have consistently shown that these documents are long, vague, and rarely read, thereby undermining the premise that consumer choice is informed. Data flows – what data is collected, how that data is used and shared, what inferences are created from that data – are complicated, and vulnerabilities emanating from these data flows are difficult to identify. In addition, consumers are often not provided an authentic choice due to the market power of an organization (e.g., web search), opt-out as the default choice (e.g., behavioral advertising), dark patterns aimed at minimizing opt-out and otherwise influencing decision-making, or consumer choices simply being ignored.

Problems with notice and choice are not limited to implementation, but are inherent in the notice and choice regime. Privacy notices serve multiple purposes – providing legal disclosures, serving as a legally-enforceable data-use contract, and providing transparency to users – so there is no notice length or level of precision in the notice that meets the needs of each of these purposes. Moreover, the time required to read these documents does not scale to the number of services with which users interact. In addition, the choice to opt-in by one data subject can impact other subjects, e.g., when the choice to disclose covers personal information from more than one subject. Finally, notice and choice inherently places all responsibility on the user to understand data practices and make an informed decision. Yet the ever-changing data flows and nuanced implications of data practices render consumers poorly-positioned to ensure organizations' data practices meet the needs of individuals or society.

User choice is not the only possible mechanism for determining whether organizations' data practices are responsible or harm individuals or society. Other industries rely on regulations to provide minimum standards as well as reporting and auditing requirements that serve to hold organizations accountable for their business practices, provide standards of appropriate behavior, and enforce societal requirements and needs on businesses. Businesses and markets retain legitimacy not only through ensuring consumers are informed and make authentic choices but also through the work of auditors, regulators, industry groups, etc. to hold organizations accountable to societal standards.

Here, we propose a new framework that describes how responsible data practices could be enforced beyond the notice and choice regime. We first identify the features necessary in a world beyond the notice and choice regime. We then propose how current and future market and regulatory incentives and stakeholder roles can help hold organizations accountable for their data practices. This includes roles internal to the organization, such as the chief

executive officer, privacy risk and compliance officers, and software engineers, among others, as well as new roles external to the organization, such as auditors and insurers. These roles work together in a "web" to help organizations be accountable for their data practices to regulators. We then outline a possible approach to define privacy standards to guide organizational data practices, potential incentives for organizations to adopt the proposed framework, as well as potential obstacles to framework implementation and how those obstacles may be avoided or overcome. Finally, we present a roadmap for how the framework implementation – beyond notice and choice – could be achieved and what documents would be necessary to support this framework.

4.4.2 Features of an effective privacy protection paradigm

In order to move beyond notice and choice, it is important to articulate the key features of an effective privacy protection paradigm – features that are supposedly being achieved through the notice and choice regime. In this section, we identify four such features – (1) transparency, (2) responsible data practices, (3) minimization of regulatory burden, and (4) legitimacy of the data market – and we outline how the current framework of the notice and choice regime falls short of these requirements. Table 2 provides an overview.

Table 2 Comparison of Privacy Protection Principles: Traditional vs. New Paradigm.

Principles of Privacy Protection	Notice and Choice	New Paradigm
Transparency	NO	YES
Responsible data practices	NO	YES
Minimization of regulatory burden	YES and NO	YES and NO
Legitimacy of the data market	NO	YES

4.4.2.1 Transparency

The current notice and choice regime aims to provide transparency over organizational practices through privacy notices. To this end, organizations share information about their data practices in a way that is supposed to be accessible and understandable to all stakeholders. To meet this standard, various stakeholders require different information presented in different ways. For the data subjects, or individuals about whom data is collected or used, we observe that transparency efforts should focus on conveying information to users in a manner that is clearly written, not overly technical, and easy to read in a reasonably short amount of time. For example, the General Data Protection Regulation (GDPR) imposes similar requirements on notices as described in Article 12. However, there are few metrics to measure whether notices meet these requirements. Visualization tools or interactive interfaces may be helpful. For regulators, technical implementation details about data practices are necessary to evaluate compliance with legal requirements.

The current notice and choice framework provides the same privacy notices to both users and regulators, leaving both parties dissatisfied. Organizations are incentivized to create long, opaque or vague, and overly broad privacy documents to limit liability; as a result, these documents are hard for users to understand and the effort necessary to read the policies does not scale to the number of services with which a user interacts. An effective

privacy protection framework would need to provide transparency of an organization's data practices by ensuring documents are (1) comprehensive, (2) specifically tailored to the needs of different stakeholders, and (3) accurate representations of actual practices.

4.4.2.2 Responsible data practices

Organizational data practices should be consistent with laws and standards, which often encode societal values. Notice and choice regimes place responsibility on the consumer to recognize risks to their personal privacy and to be the sole decision maker about whether that risk is acceptable. Consumer choice acts as the sole force to ensure organizations' data practices are responsible and meet the needs and values of individuals and society. However, the shortcomings of notice and choice – including unreadable policies, incomprehensible implications of data practices, unscalable user burden, and limited choices – ensure that, in practice, users cannot always reasonably choose to use only services with responsible data practices. This often results in adoption of services that are inconsistent with user values, such as those that repurpose user data. To be effective, a privacy protection framework must employ incentives that go beyond the status quo. This includes new actor roles to hold organizations accountable and ensure responsible data practices, while incorporating robust enforcement mechanisms to penalize practices that violate legal standards or social values without relying only on consumer choice.

4.4.2.3 Minimization of regulatory burden

For a regulatory approach to be effective, enforcement mechanisms must be practical and enforceable. Regulators have finite resources that impose practical bounds on their actions; if misbehavior is less likely to be detected, organizations have less incentive to comply. The notice and choice regime minimizes regulatory burden by placing the primary responsibility for ensuring responsible data practices on the data subjects making the "correct" choice, which is criticized as increasingly being an "illusion of choice." On the other hand, the lack of transparency in current notices places a significant burden on regulators, who must investigate current practices without access to detailed internal information or technical details. Rather than rely on regulators to conduct random investigations, a new framework that requires organizations to disclose risks to privacy and non-compliance can guide regulators toward which organizations are likely to be non-compliant. The consequence of such disclosures can further motivate organizations to be more responsible. An effective privacy protection framework should provide incentives, structure, and access that enable effective external regulation without imposing undue regulatory burden.

4.4.2.4 Legitimacy of the data market

A successful privacy framework should engender trust in the marketplace, ensuring that any data transaction is made without fraud, manipulation, or deception. In the notice and choice regime, fully-informed choice presumes accurate information being received by data subjects and ensures individual autonomy, thereby legitimizing transactions over personal data. The broader market's legitimacy is thus determined through the summation of the legitimacy of these transactions. However, this regime assumes data subjects have the time, technical knowledge, and understanding of the broader technology ecosystem to make sense of privacy notices. Research has consistently shown that these requirements are not met, rendering this regime ineffective. An effective privacy protection paradigm should ensure legitimacy of the data market through a multi-faceted approach.

4.4.3 Research directions

In the light of these challenges and key features of an effective privacy protection paradigm, we identify directions for future research, shown in Table 3.

Table 3 Research directions.

Topic	Immediate	Medium-term (3 yrs)	Long-term (10 yrs)
Laying out the privacy protection paradigm *(Econ, Policy, Law, CS, Business)*	Identifying features and expectations from an effective privacy protection paradigm. Enumerating the different market forces, stakeholder roles and responsibilities that may be operationalized to motivate responsible data markets and compliance within them.	Providing specifications and examples of documents needed to support the paradigm. *(HCI, Business, Law)* Drafting regulation in support of the new paradigm.	
Defining responsible data practices *(Econ, Policy, Law, CS, Business)*	Developing models and frameworks for defining responsible data practices. How do we evaluate and assess which data practices are consistent with societal values in different contexts?	Develop mechanisms for assessing fines and incentives within the proposed frameworks. Developing mechanisms to evaluate the effectiveness of different frameworks for responsible data practices.	Developing mechanisms for transitioning frameworks to practice and assessing their performance. How do we assess harms and gains made from violating proposed frameworks?
Transparency needs and responsibilities *(Econ, Policy, Law, CS, Business)*	What are the transparency needs of different stakeholders? Does existing HCI research already suggest the best ways to address these needs? What documents are needed to provide effective transparency, and what should these documents look like?	Additional HCI research to address dark patterns in transparency documents and other transparency gaps/challenges.	
Professional obligations *(Econ, Policy, Law, CS, Business)*	Identify the professional obligations of a licensed CS engineer. Design document formats that professionalized engineers can use to communicate technical details and risks to internal privacy risk officers.	Design the conditions under which a licensed computer scientist is required.	Developing licensing bodies and mechanisms.
Whistleblower Employees *(Policy, Law, Business)*	Research when whistleblowers are needed based on regulatory and reporting requirements.	Identify the conditions and protections for whistleblowers in tech.	

Topic	Immediate	Medium-term (3 yrs)	Long-term (10 yrs)
Audits *(Econ, Policy, Law, CS, Business)*	Design auditing requirements for government contracting and special industries. Design documents that outline specific auditing requirements.	Expand the conditions requiring data audits including for privacy risk reporting for SEC.	Expand conditions requiring data audits to include organizations with individualized data.
SEC/Shareholder reporting *(Econ, Policy, Law, Business)*	Increase risk of high fines for privacy violations through state and federal regulations/policies. Design document formats for communicating organization behavior to regulators to enable meaningful compliance auditing.	Design reporting obligations for privacy risk assessments for publicly traded companies to SEC. Design the role of a privacy risk officer with obligations to report privacy risk based on potential fines and organizations' data practices.	
Insurance *(Econ, Policy, Law, Business)*	Increase risk of high fines for privacy violations (see above).	Design insurance market for organizations wishing to mitigate privacy risk.	

4.5 Cross-Stakeholder Interaction

Jose M. del Alamo (Polytechnic University of Madrid, ES), Soheil Human (Wirtschaftsuniversität Wien, AT), Konrad Kollnig (Maastricht University, NL), Daniel Smullen (CableLabs – Louisville, US), and Kami Vaniea (University of Waterloo, CA)

4.5.1 Challenges

Without noticing it much, we are all subject to a wealth of different privacy documents every day. These include privacy policies, FAQs, and the text shown in consent pop-ups of the various online services that we all use. The breadth of document types is impressive, encompassing terms of service, cookie policies, privacy disclosures, engineering specifications, and more. These documents also include legal statutes, implementing and delegated acts, and regulatory guidelines that underpin data processing. Often, laws from multiple countries apply to a single data processing operation of the same company, given the global scale of the Internet.

Instead of end-users (i.e., the public writ large), legal professionals – foremost lawyers and judges – are those who are currently most empowered by the broad share of currently used privacy document types. Perhaps the most common type of privacy document, privacy policies, in theory are to inform end-users about how their data is used. However, these documents are drafted mainly for lawyers and judges, thereby leaving end-users struggling to understand legal language in order to understand what happens with their data. Similarly, governments are those who enact privacy laws but legal documents are written in a language specific to law, which can be challenging for engineers to understand in terms of computer code and/or system implementation needs. Similarly, these documents are extremely difficult for end-users to interpret, even though they are the data subjects that the documents are intended to enshrine rights to.

While the list of stakeholders has yet to be fully enumerated, and identifying the broad set of documents that relate to them is an open research problem, we can already conclude some basic facts about all privacy documents. Among all stakeholders, from legal to end-user to engineer and beyond, the main purpose of privacy documents is the communication of privacy concepts from one stakeholder group to another. This underscores that the study of privacy documents is the study of communication between pairs of stakeholder groups, whose individual and collective needs must be considered when trying to make any such communication work. In practice, privacy documents often serve multiple stakeholder groups, with each of their unique needs. Too often, privacy documents are designed around a narrow group of stakeholders but used in a way that requires stakeholders outside of this scope to consume them.

There are many other stakeholders beyond data subjects and legal professionals, such as engineers in data-processing companies or competitors, members of civil society organizations, and parents of children who are subject to data processing. These, too, face the general challenge that they are not empowered by the privacy documents that other stakeholders use to communicate with them. The consequences are numerous; stakeholders are inadequately informed by the privacy documents they consume. The documents are at the wrong level of abstraction for their needs. The documents are intended to be understood in a way that may be familiar to the producer but is unfamiliar to the document consumer. In general, the misalignment of documents with their stakeholders exhibits many problems related to their application and suitability for use in numerous real-world scenarios that deserve further study. As we see in our previously cited examples, such as lawyers communicating with end-users (via privacy policies), there is an obvious but poorly understood mismatch in the communication these documents facilitate between the various pairs of stakeholders (i.e., document producers and consumers) and their intent. The result are a lack of transparency, which results in information asymmetry, and all the downstream problems that arise from under- or over-specification and misunderstanding. We view the overarching challenge in this space as one that arises from a lack of shared expertise between stakeholder groups. Without an interdisciplinary effort to improve communication between these groups, each stakeholder runs the risk of ineffective communication.

Figure 1 provides an abstract vignette, notionally illustrating how privacy documents serve as a means of communication between a few pairs of imaginary stakeholders. This illustration also captures the common scenario we see in the real world – that the same document is often used to facilitate the communication between different stakeholders in a way which is well-intentioned but falls short of achieving its intent. When companies use the same documents (i.e., privacy policies) to communicate with both regulators and users, the mismatch between documents intended for companies and regulators to communicate results in a communication failure.

Better communication arises from interdisciplinary work, such as where experts in law and technology work together to more holistically study the problems in this space and develop better solutions to address them. These solutions often revolve around the ideal state of having prescribed formats, structures, or schemata for communicating privacy information and concepts. Prominent examples are the Platform for Privacy Preferences Project (P3P) and Privacy Nutrition Labels, in theory enabling rapid or even automated communication. However, these solutions have rarely been adopted into practice or sustained so that they would meaningfully empower the involved stakeholders – end-users in particular. This situation reveals a key challenge: Too often, one stakeholder group (e.g., computer science academia, advertising industry, privacy regulators, browser developers) comes up

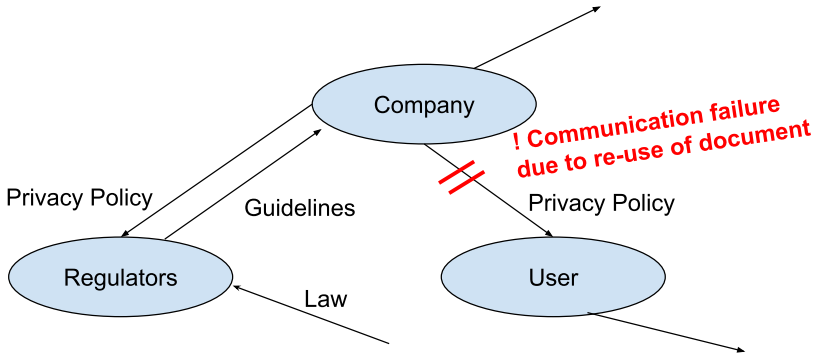


Figure 1 Notional example of how privacy documents serve as a means for communication between a few different imaginary stakeholders. Ineffective communication is common.

with solutions that generalize well within their own stakeholder group or a narrow slice of other groups but do not adequately address the needs of all stakeholders involved. Thus, work towards the adoption of those solutions by others falls short of meeting that goal. They can fail to take into account more pluralistic approaches and miss opportunities for better tailored approaches intended for specific stakeholder pairs.

Addressing a wide range of stakeholders requires interdisciplinary research, which everyone is in favor of in principle. But in practice incentives are missing: access to funding, venues for publication, means for long-term development of tooling, recognition in academic promotion, and so on. As a result, research ends up focusing on narrow parts of the inter-stakeholder relationships, where it aligns with the existing incentives.

In sum, we identify three main challenges:

- Stakeholder imbalance in needs around privacy documents in relation to motivation, information needs, empowerment, specificity, and clarity.
- Researcher needs to conduct interdisciplinary research that aims to connect several disciplines to understand problems and find solutions.
- Robust implementation strategies for how solutions will be developed, adopted, and enforced, as well as an understanding of how the strategies might be resisted.

4.5.2 Key research questions

- What mechanisms can be developed to ensure that theoretical research on privacy documents is translated into practical, industry-wide applications?
- How can research outputs in the design and analysis of privacy documents be made more accessible and relevant to policymakers and industry practitioners?
- What are the barriers to adopting academic innovations in privacy policy tools and practices within organizational settings?

4.5.3 Research directions

Addressing these challenges requires a multifaceted strategy. It is essential to enhance the dialogue between researchers and practitioners to ensure that research outputs are designed with practical constraints and needs in mind. Developing flexible, scalable solutions that can be easily integrated into existing systems and processes is crucial. Furthermore, fostering a regulatory environment that supports and promotes the enforcement of these novel solutions can provide the necessary impetus for their broader adoption.

4.5.3.1 Mapping the ecosystem of privacy management

A comprehensive understanding of the complex network of stakeholders, their interrelationships, and the channels through which communication and enforcement occur is vital. This involves mapping out the ecosystem of privacy management, identifying how different stakeholders interact, and recognizing the influence they exert on each other. Such an understanding can help in tailoring solutions that are not only technically sound and legally compliant but also socially and organizationally feasible.

4.5.3.2 Motivating privacy for stakeholders

Educating all stakeholders about the potential benefits and long-term gains of implementing advanced privacy solutions will be key to overcoming resistance and achieving widespread acceptance. This holistic approach is paramount to successfully translating privacy research into effective, enduring practices.

4.5.3.3 Communicating challenges in domain-specific language

Existing research already aims to understand the needs of different stakeholders and stakeholder pairs. Extensive research in the social sciences, for example, looks at the needs of vulnerable groups. Similarly, mapping the different actors in privacy enforcement is likely studied by many researchers. The challenge is that this work is currently presented in ways that are appropriate for the domains it was conducted in. More effort is needed to enable cross-domain presentation of research in ways that are accessible and help researchers better understand the challenges. In particular, survey or summarization type research would be quite valuable to provide alternative presentations of that are meant to be consumed by other disciplines.

4.5.3.4 Researcher incentivization and support

Addressing the lack of incentives and support for researchers in interdisciplinary fields requires structural and cultural changes within academic and research institutions, as well as funding bodies. Some solutions and ideas to move towards that end are:

- *Promote funding.* Both the National Science Foundation (NSF) and Horizon Europe (HEU) have initiatives to fund interdisciplinary research and promote collaboration, yet more targeted (or open-ended) programs can encourage collaboration across disciplines. In Europe, COST Actions and HEU Marie Curie networks can be helpful instruments to that end. The former can be used to create professional societies/chapters or online platforms to connect interdisciplinary researchers. The latter supports developing programs that train students and early-career researchers in methods and theories from multiple disciplines.
- *Revise academic incentive structures.* The metrics for academic advancement should include recognition of interdisciplinary work, e.g., by considering DORA-like criteria in assessing scientific quality and promotion. Identifying and sharing (or setting up if not available) reputed venues for interdisciplinary publications and/or cross-disciplinary outreach will enable this recognition by peers, further supporting networking and collaboration.

Participants

- Noah Apthorpe
Colgate University – Hamilton, US
- Eleanor Birrell
Pomona College – Claremont, US
- Travis Breaux
Carnegie Mellon University – Pittsburgh, US
- Kai-Wei Chang
UCLA, US
- Jose M. del Alamo
Polytechnic University of Madrid, ES
- Rinku Dewri
University of Denver, US
- Nico Ebert
ZHAW – Winterthur, CH
- Simone Fischer-Hübner
Karlstad University, SE
- Sepideh Ghanavati
University of Maine, US
- Henry Hosseini
Universität Münster, DE & Westfälische Hochschule – Gelsenkirchen, DE
- Soheil Human
Wirtschaftsuniversität Wien, AT
- Agnieszka Kitkowska
Jönköping University, SE
- Konrad Kollnig
Maastricht University, NL
- Kirsten Martin
University of Notre Dame, US
- Jelena Mitrovic
Universität Passau, DE & Institute for Artificial Intelligence R&D of Serbia – Novi Sad, RS
- Rishab Nithyanand
University of Iowa – Iowa City, US
- Shidong Pan
Australian National University – Acton, AU
- Sarah Radway
Harvard University – Allston, US
- Tim Samples
University of Georgia, US
- Florian Schaub
University of Michigan – Ann Arbor, US
- Yan Shvartzshnaider
York University – Toronto, CA
- Daniel Smullen
CableLabs – Louisville, US
- Peter Story
Clark University – Worcester, US
- Emma Tosch
Northeastern University – Boston, US
- Christine Utz
Radboud University Nijmegen, NL
- Kami Vaniea
University of Waterloo, CA
- Isabel Wagner
Universität Basel, CH
- Shomir Wilson
Pennsylvania State University – University Park, US
- Maximiliane Windl
LMU München, DE
- Lu Xian
University of Michigan – Ann Arbor, US
- Tianyang Zhao
Pennsylvania State University – University Park, US

Report from Dagstuhl Seminar 25022

Towards a Multidisciplinary Vision for Culturally Inclusive Generative AI

Asia Biega*[1], Georgina Born*[2], Fernando Diaz*[3], Mary L. Gray*[4], and Rida Qadri*[5]

1 **MPI-SP – Bochum, DE.** asia.biega@acm.org
2 **University College London, GB.** g.born@ucl.ac.uk
3 **Carnegie Mellon University – Pittsburgh, US.** diazf@acm.org
4 **Microsoft New England R&D Center – Cambridge, US.** mlg@microsoft.com
5 **Google – San Francisco, US.** ridaqadri@google.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25022 "Towards a Multidisciplinary Vision for Culturally Inclusive Generative AI". The gathering focused on questions raised by the rapid deployment of Generative AI systems and their integration into global systems of cultural communication, consumption, and production. As these technologies shape our cultures, we urgently need conceptual foundations for investigating the cultural inclusivity of generative AI pipelines (from data collection, to model development and deployment, to evaluation), as well as methods to study the varying societal and cultural impacts of generative AI.

This Dagstuhl Seminar convened scholars and practitioners from computer science, social sciences, the tech industry, and creative industries to discuss the cultural implications of generative AI and find paths toward building generative AI that can be responsive to the diverse needs of individuals, groups, and societies around the world. Together, seminar participants began the challenging but necessary work of building shared language and frameworks for reshaping the technical and social architectures of generative AI.

The seminar was structured along three main dimensions for interdisciplinary discussions:

- Examining the cultural values being currently centered in generative AI.
- Studying the possibilities and risks of encoding cultural knowledge into generative AI technologies.
- Understanding the cultural impact of these technologies.

We succeeded in building an expert network committed to understanding and designing a culturally-attuned generative AI and to lay the foundation for an interdisciplinary research and practice agenda on global inclusion and generative AI.

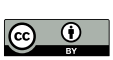
Seminar January 6–9, 2025 – https://www.dagstuhl.de/25022

2012 ACM Subject Classification Applied computing; Computing methodologies; Human-centered computing; Information systems

Keywords and phrases creativity, cultural inclusion, generative artificial intelligence, global south, social impact of ai

Digital Object Identifier 10.4230/DagRep.15.1.33

* Editor / Organizer



Towards a Multidisciplinary Vision for Culturally Inclusive Generative AI, *Dagstuhl Reports*, Vol. 15, Issue 1, pp. 33–49
Editors: Asia Biega, Georgina Born, Fernando Diaz, Mary L. Gray, and Rida Qadri

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

1 Executive Summary

Rida Qadri (Google – San Francisco, US)
Asia Biega (MPI-SP – Bochum, DE)
Georgina Born (University College London, GB)
Fernando Diaz (Carnegie Mellon University – Pittsburgh, US)
Mary L. Gray (Microsoft New England R&D Center – Cambridge, US)

Motivation

Recent years have seen rapid development and widespread adoption of generative AI systems that algorithmically model human creativity and decision-making. In particular, this technological shift has profound implications for how cultural artifacts like music, news, literature, and film are produced and consumed, raising concerns about the potential cultural implications of this technology. At the same time, these technologies are displaying Western-centrism in AI training and evaluation data, definitions of "success", and evaluation methods. As a result, generative AI technologies, while arguably improving their reliability for basic output of sensible prose and images, have a recognizable pattern of failing to generate norms and values representative and inclusive of non-Western perspectives. For example, recent research and media reports have found that models are less competent at generating culturally significant material outside of a Western point of view, frequently omitting non-Western cultural knowledge from outputs, and perpetuating Western stereotypes in generated output. Addressing these failures and their broader impact is crucial to prevent globally-launched generative AI tools from becoming vehicles for reinforcing Western-centric cultural norms and values, production and distribution methods, and in these ways further exacerbating global inequities.

The urgent need for a seminar on these topics was highlighted in the first-of-its-kind 2022 NeurIPS workshop on "AI and Culture" that brought together researchers from computer science, the humanities, and the social sciences at the premier conference of AI researchers and practitioners. At this workshop, emergent conversations pointed to how building culturally sensitive, responsive, and accountable AI systems will require researchers and engineers to include diverse disciplinary voices, community expertise, and cultural knowledge in AI research and development. Such efforts to recognize and incorporate myriad cultural contexts into AI systems are often siloed within disciplines and, as a result, are disjointed and limited in their impact on technological design. In particular, there are no cohesive frameworks to help researchers fold nuanced cultural analyses and situated knowledge into generative AI models. There is therefore a critical, currently unmet need to break down disciplinary silos and create coherent interdisciplinary conceptual foundations for novel, culturally-sensitive generative AI research and practices. The most promising areas in need of interdisciplinary collaborative research include: 1) new approaches to data collection; 2) interdisciplinary frameworks and methods for model development and deployment; and 3) new techniques that integrate and distinguish the value of qualitative and computational approaches to evaluation. We also see the need to develop new interdisciplinary methods, crossing between qualitative and quantitative approaches, to study the societal impacts of generative AI. As generative AI research is currently confined mainly to industry-academic collaborations, we further aim to broaden the contextual and institutional perspectives brought to these challenges beyond academia to include voices from civil society and impacted communities – and this was also a goal achieved in the seminar.

Program

The seminar lasted 2.5 days. As our goal was to create an interdisciplinary space for discussion, we had 28 participants with backgrounds in multiple disciplines and sectors. Participants included experts in computer science, data science, machine learning (ML), information retrieval (IR), natural language processing (NLP), human-computer interaction (HCI), responsible artificial intelligence (AI), social computing, critical data studies, music and ethnomusicology, anthropology, history, political philosophy, science and technology studies (STS), media studies, communication, and architecture. The seminar also included contributions from filmmakers and the creative industry. The participant pool reflected the broad spectrum of perspectives and expertise on language, culture, and cultural production, necessary to advance the dialogue on AI and culture.

To encourage participants to come to the seminar prepared with preliminary reflections on the topic, we asked them to complete a round of preparatory work two months before the seminar. This consisted of sharing a paper that participants had written or found fruitful in their current work in order to introduce themselves to the rest of the group and explain their way of approaching questions of AI and culture. We also asked each participant to reflect on a series of questions: 1) How are you thinking about the term "culture" in the context of artificial intelligence? 2) What is a provocation or critical question you would like to share regarding the intersection of AI and culture? And, 3) Where would you like to see the field of AI and culture head next?

The first day of the seminar was dedicated to sparking discussion and allowing participants to get to know each other's perspectives on the seminar topic. Recognizing that most of the invited scholars do not regularly cross paths at a single-disciplinary home conference, the first exercise of the day was a series of "speed dating" rounds. Participants rotated through ten-minute introductory conversations with at least three other participants. We asked participants to share basic information about their disciplinary training and home institutions. Then they added background on what they hoped to gain in terms of a deeper understanding of the interdisciplinary challenges and opportunities in the emerging field of generative AI and cultural diversity. During Round 2, participants shared the next project they were working on or their dream project in this space. In Round 3, they discussed examples of AI failures that illustrated their thinking on the seminar topic. These discussions helped to identify the first key areas for the advancement of the field.

Once participants had a sense of the breadth and depth of expertise in the room, we shifted to the first substantive programming component. This took the form of three panels each with three speakers, with each speaker offering 5-minute "firestarter" provocations, followed by an open group discussion. The firestarter presentations were followed by a short, individual reflective writing session, where participants could document their questions and reactions to the discussions, and contribute to a collective note-taking document. The nine speakers were selected to give firestarter talks on the basis of the submitted preparatory work. The organizers conducted thematic coding of the received documents ahead of the seminar in order to assemble the panels, with the aim of putting participants into multi- and interdisciplinary dialogue early on in the seminar.

On the second day of the seminar, participants were invited to collectively come up with themes they wanted to discuss further. They then broke into small group discussions on the chosen themes. The small group discussions were followed by shareout sessions, followed by the generation of provocations, and the genesis of potential future collaborative projects among participants. As noted above, we used the themes and points of friction from Day 1's Firestarter discussions and individual reflections to brainstorm and then thematically code

the questions and clusters of discussion that had the most consensus. After spending the morning consolidating themes, we identified and converged on three clusters for small group discussion. The themes were articulated as: Discussion Cluster 1: Power, Future, History; Discussion Cluster 2: Interdisciplinarity in Computer Science Cultures; and Discussion Cluster 3: Culture Encodability. Each Discussion Cluster is described below.

On the third day of the seminar, participants discussed the next steps in small groups.

Outcomes / Planned outcomes

This seminar fostered a critical reflection on the development of culturally inclusive AI, highlighting the rare opportunity for interdisciplinary learning, and generating a profound sense of urgency and clarity regarding the challenges. Participants formed working groups around three key outcomes: an agenda-setting document, a "meta-metadata" project, and a research project on integrating qualitative and quantitative evaluation methods.

Agenda-Setting Project This initiative aims to establish a shared, nuanced understanding of AI, culture, and technology, moving beyond simplistic definitions. The group will produce a document for funders like the NSF, articulating the challenges and relevance of culturally inclusive AI. This document will influence funding priorities and foster interdisciplinary research, serving as a foundation for broader engagement with policymakers and the public.

Meta-Metadata Project This research project focuses on developing and implementing new approaches to metadata creation and management, fostering culturally rich datasets through open-source, collaborative models. Key planned outcomes include a course and hackathon exploring nuanced metadata encoding, and the creation of a network of scholars dedicated to this work. The project also explores leveraging existing platforms like Wikimedia to host and manage detailed, culturally diverse metadata, addressing challenges like image metadata and incentivizing scholarly participation.

Project on Integrating Qualitative and Quantitative Methods for AI Evaluations This future project addresses the critical need for robust methodologies that integrate qualitative and quantitative data in AI evaluation. It aims to develop frameworks that translate qualitative insights into concrete algorithmic interventions without losing critical nuances. The research will explore methods like "fictions" or imagined scenarios to anticipate potential consequences, and guide development, moving beyond the limitations of current practices that rely on small user groups or subjective judgments.

Beyond these tangible projects, the seminar achieved a significant shift in perspectives. Computer scientists gained a deeper appreciation for the complexities of culture, while social scientists and humanities scholars refined their critiques through a clearer understanding of AI's potential. This cross-disciplinary dialogue led to a richer understanding of the multi-layered relationship between AI and culture, moving beyond simplified encodings and benchmarks. Participants also valued the seminar's global representation, moving beyond US/EU-centric viewpoints. The seminar generated significant momentum, energizing participants and sparking new collaborative research directions. As one computer scientist noted, "The questions I came in with are very different from the questions I'm leaving with... I find that the questions I leave with are much richer – and harder." Similarly, a social scientist expressed, "As a non-technical person the seminar was incredibly insightful to better understand what the state of the art currently is, what the possibilities and limitations for culturally sensitive interventions in these systems may be."

Participants consistently highlighted the need for a second iteration of the seminar, emphasizing the value of continued multidisciplinary spaces for collaboration. They left with a richer set of concerns and vocabularies, anticipating that this assemblage would transform individual disciplinary research and lead to numerous joint collaborations. The seminar was described as "creatively fortifying and vitalizing," creating meaningful connections and inspiring participants to push forward in the pursuit of culturally inclusive AI.

2 Table of Contents

3 Overview of Talks

3.1 Firestarters: Initial Areas for Exploration (25022)

Rida Qadri (Google – San Francisco, US), Asia Biega (MPI-SP – Bochum, DE), Georgina Born (University College London, GB), Fernando Diaz (Carnegie Mellon University – Pittsburgh, US), and Mary L. Gray (Microsoft New England R&D Center – Cambridge, US)

Three high-level themes emerged after we moved to individual reflection and a last round of open discussion, setting the stage for Day 2's Clustering exercise. Specifically, we ended our day noting the following key areas for exploration:

- The challenges of defining culture are multifaceted, involving different definitions and disciplinary lenses. There are significant gaps in what is being represented, and understanding what culture can achieve beyond Responsible AI ethics framings is crucial.
- Encoding culture within AI systems presents its own set of challenges, including the use of various computational methods to incorporate cultural knowledge and their potential consequences.
- The role of metadata is complex, and integrating both quantitative and qualitative perspectives is essential yet challenging.
- Institutional aspects and interdisciplinarity play a significant role in the cultures of AI production. There is a need for alternative imaginaries of technology that go beyond the corporate inclusion of data.
- Building collaborative teams that include diverse perspectives and experiences is vital, and fostering interdisciplinarity is key to advancing the field.
- The headwinds that work against multidisciplinary approaches to culturally-inclusive AI are exacerbated by the absence of regulatory frameworks and cultural norms that could foster synergies and accountability across academic and industry-based AI research and development settings.
- The challenges of definitions of culture (different definitions and disciplinary lenses, what is not being represented, what culture can get us beyond Responsible AI ethics framings).
- Encoding culture (different computational methods to include cultural knowledge in AI – and their consequences, the complex role of metadata, and challenges of integrating quantitative and qualitative perspectives).
- Institutional aspects and interdisciplinarity (cultures of AI production and alternative imaginaries of tech, AI beyond corporate inclusion of data, building collaborative teams, interdisciplinarity).

3.2 Panel Discussion 1: Definitions of Culture (25022)

Rida Qadri (Google – San Francisco, US), Hal Daumé III (University of Maryland – College Park, US), Tarleton Gillespie (Microsoft New England R&D Center – Cambridge, US), and Molly Steenson (American Swedish Insitute – Minneapolis, US & Carnegie Mellon University – Pittsburgh, US)

Panel Discussion 1 at the Dagstuhl Seminar focused on the theme of "Definitions of Culture" and was presented by Tarleton Gillespie, Hal Daume, and Molly Wright Steenson.

The session began with Tarleton Gillespie opening with a quote from Raymond Williams, highlighting the complexity of defining culture. Gillespie emphasized the importance of representation and the lived practices of stakeholders, including designers and users. He raised questions about how cultural values are inscribed in tools and the biases that emerge over time.

Hal Daume's presentation focused on the gaps between community knowledge and computational knowledge. He discussed the challenges of measuring culture and the limitations of current AI systems in understanding diverse cultural contexts. Daume highlighted the mismatch between the knowledge of individuals and communities and the knowledge embedded in AI systems, using examples such as sign language and African American linguistic communities. He questioned whether computer science is open to expanding its understanding of culture beyond quantifiable metrics.

Molly Wright Steenson's contribution focused on the cyclical nature of how industries manage crises related to ethics and safety. She reflected on the ethical crisis in Responsible AI (RAI) in 2018 and what lessons could be learned to rethink the framework. Steenson also discussed the importance of considering cultural practices and norms in the development of AI systems and the potential for cultural imposition by organizational structures. She emphasized the need for a hybrid methodology that integrates qualitative and quantitative approaches to better understand and model cultural norms.

The key takeaways from Firestarter 1 included the recognition of the complexities and tensions in defining and measuring culture within AI systems. The presenters highlighted the importance of expanding the understanding of culture in computer science, moving beyond mere quantification to include qualitative insights. They also highlighted the need for interdisciplinary collaboration and the inclusion of diverse cultural perspectives in AI development.

3.3 Panel Discussion 2: Encoding Culture (25022)

Rida Qadri (Google – San Francisco, US), Mary L. Gray (Microsoft New England R&D Center – Cambridge, US), Huma Gupta (MIT – Cambridge, US), Emanuel Moss (Intel – Santa Clara, US), and Alice Oh (KAIST – Daejeon, KR)

Panel Discussion 2 focused on the theme of "Encoding Culture" and was presented by Huma Gupta, Alice Oh, and Emanuel Moss.

Huma Gupta

Huma Gupta's presentation centered on the concept of the "Library of Missing Metadata," inspired by Mimi Onuoha's work. Gupta explored the idea of adding meta-metadata to support changing interpretations of artifacts and the complexities of digital architectures. She discussed the aggregation of terms and the legacies of taxonomies, suggesting that meta-metadata could introduce friction and complexity to visually prompt disruption of what counts as culture. Gupta also reflected on the challenges of encoding culture and the potential for cultural imposition by organizational structures.

Alice Oh

The session began with Alice Oh discussing her expertise in building large language models (LLMs) and the challenges of creating benchmarks for cultural competence. She emphasized the importance of considering a mix of well-represented and under-represented cultures and highlighted the difficulties in defining what should be included in these benchmarks. Oh also pointed out the presuppositions embedded in questions and the need for application scenarios to create effective LLMs.

Oh specifically discussed the BLEnD Dataset which represents their recent effort to evaluate the cultural commonsense knowledge of large language models (LLMs) across 13 languages and 16 regions. Native speakers collaboratively created a common set of questions, translated them into their languages, and gathered responses from other native speakers. The evaluation of the LLMs on this carefully crafted dataset highlighted serious limitations in LLMs: they struggle to perform well in understanding and representing languages and cultures outside of a few dominant ones.

However, the project also raised deeper questions about methodology and objectives. For example, what exactly is "culture," and how can we ensure that questions posed to annotators avoid embedding cultural presuppositions? Determining the "ground truth" for answers further complicates matters, as cultural identity is complex and multifaceted. A Korean annotator, for instance, might draw on their heritage, personal experiences, and exposure to other cultures, such as life in the U.S. or work in a global field like computer science. LLMs must be designed to navigate such complexities by recognizing the existence of multiple perspectives and acknowledging that some questions or answers can be sensitive or offensive. This means we need careful grounding of the evaluation process, defining cultural knowledge in concrete terms, and considering real-world usage scenarios where LLMs must perform reliably and inclusively.

Emanuel Moss

Emanuel Moss contributed to the discussion by emphasizing that culture cannot be encompassed by any individual or described by simple rules. He described AI as an artifact of culture and defined culture as a set of shared conceptions expressed through symbolic forms. Moss raised questions about the possibility of benchmarking and encoding culture, the relational and collective aspects of culture, and the potential harms of trying to benchmark culture. He also highlighted the importance of considering the processes of cultural production and the risks of encoding culture into corporate databases and models.

The key takeaways from Firestarter 2 included the recognition of the complexities and challenges in encoding culture within AI systems. The presenters illuminated the importance of considering diverse cultural perspectives and the potential harms of misrepresentation and exclusion.

3.4 Panel Discussion 3: Institutional Reflections and Collaborations (25022)

Rida Qadri (Google – San Francisco, US), Naveen Bagalkot (Manipal Academy of Higher Education – Bangalore, IN), Catherine d'Ignazio (MIT – Cambridge, US), and Sara Hooker (Cohere For AI – Toronto, CA)

The seminar then shifted to its third and final panel focused on the theme of "Institutional Reflections and Collaborations" that was presented by Catherine D'Ignazio, Sara Hooker, and Naveen Bagalkot.

Catherine D'Ignazio

The session began with Catherine D'Ignazio discussing her multi-year project working with activists to do participatory work at every stage of AI development. She emphasized the importance of privileging subjugated knowledges and questioned the current organization of resources in the AI ecosystem. D'Ignazio highlighted the need for sustainable and just resources, envisioning alternative initiatives that do not rely on corporate structures.

Sara Hooker

Sara Hooker, who leads Cohere for AI, discussed the importance of surplus and excess in driving innovation. She reflected on the history of scientific breakthroughs and the need for open science collaboration across many companies. Hooker raised concerns about the marginalization of academia and the underrepresentation of researchers from the Majority World. She emphasized the need for alternative, sustainable, and just approaches to AI development that support participation and inclusivity.

Naveen Bagalkot

Naveen Bagalkot contributed the last point of view to the discussion by sharing a narrative about the futures of AI. He highlighted the importance of centering the processes of cultural production and considering how technologies and interactions are result of these processes. Bagalkot emphasized the need for alternative imaginaries of technology and the importance of building collaborative teams that include diverse perspectives. He also discussed the challenges of changing research culture and the need for new funding structures that support interdisciplinary collaboration.

The key takeaways from Firestarter 3 included the recognition of the need for alternative approaches to AI development that prioritize sustainability, justice, and inclusivity. The presenters discussed the importance of participatory methods, open science collaboration, and the inclusion of diverse cultural perspectives. They also highlighted the challenges of the current corporate-dominated AI ecosystem and the need for new funding structures and research cultures that support interdisciplinary and inclusive innovation.

Each Firestarter session concluded with a call to rethink how cultural values are embedded in AI tools and to ensure that these tools are sensitive to the cultural contexts in which they operate.

4 Working groups

4.1 Working Group 1: Power, Future, History (25022)

Rida Qadri (Google – San Francisco, US), Virgilio Almeida (Federal University of Minas Gerais – Belo Horizonte, BR), Naveen Bagalkot (Manipal Academy of Higher Education – Bangalore, IN), Georgina Born (University College London, GB), Anita Say Chan (University of Illinois at Urbana Champaign, US), Hal Daumé III (University of Maryland – College Park, US), Catherine d'Ignazio (MIT – Cambridge, US), Giovanna Fontenelle (Wikimedia – Sao Paulo, BR), Tarleton Gillespie (Microsoft New England R&D Center – Cambridge, US), Darci Sprengel (King's College – London, GB), Molly Steenson (American Swedish Insitute – Minneapolis, US & Carnegie Mellon University – Pittsburgh, US), and Harini Suresh (Brown University – Providence, US)

Working Group 1 focused on the intricate interplay between power, future, and history in the context of AI and cultural inclusivity. This cluster, involving participants that represented a cross-section of institutional and disciplinary perspectives (Giovanna, Virgilio, Naveen, Harini, Catherine, Tarleton, Georgina, Molly, Darci, Anita, and Hal), delved into the visionary aspects of AI development and its implications for society. The discussions revolved around the vision, architecture, governance, and barriers to creating a network of shared infrastructures that foster alternative imaginations and inclusivity.

One of the central themes was the importance of situating AI development within the current socio-political context, including the rise of populism and authoritarianism. The participants emphasize the need to link AI initiatives to alternative histories, such as the cybernetic turn in India and socialism in Chile, to draw lessons from public institutions like libraries, universities, and archives. They discussed the concept of "defensive localism," which involves creating urgent coalitions against AI authoritarian surveillance without requiring absolute political unity. This approach contrasts with "prospective, future-oriented place-based localisms," which focus on long-term engagement with local politics to achieve justice and inclusivity.

The cluster also explored the idea of "open AI" and the challenges associated with it. While openness is seen as crucial to avoid the concentration of power, there are critiques of the concept, such as the limitations of open systems that require significant compute and technical knowledge. The participants discussed the potential of local communities to build and train their own models, considering the trade-offs of cost and quality. They highlighted the importance of a decentralized and federated structure that links smaller, local models to avoid dependency on global models. Such an approach could create an alternative ecosystem that aligns with the public good and the common good.

Another key takeaway was the need to address the material and affective demands of participation in AI development. The participants emphasized the importance of ensuring that data contributors' livelihoods and incomes improve as a result of their participation. They discussed the challenges of engaging local communities in AI projects and the need for new kinds of education that speak to needs outside of commercial tech. The discussions also touched on the history of alternative ideologies in countries like Brazil and the need to create conditions for inclusivity that represent pluralism.

The cluster concluded with a call to formulate research questions that address the uncertainties and challenges identified. These questions could include how interdisciplinary collaboration can effectively identify and address the ethical and social risks of language models, how small language models can contribute to responsible innovation, and how to design decentralized infrastructure architectures that enable users to choose how they share and distribute their data and models. The participants also highlighted the need to pluralize the political economy of technology and reimagine futures through diverse cultural imaginaries.

In summary, the thematic cluster on power, future, and history emphasized the importance of situating AI development within a broader socio-political context, addressing the challenges of openness and decentralization, ensuring the material and affective demands of participation, and formulating research questions that guide future interdisciplinary collaboration.

4.2 Working Group 2: Interdisciplinarity / CS cultures (25022)

Rida Qadri (Google – San Francisco, US), Asia Biega (MPI-SP – Bochum, DE), Tobias Blanke (University of Amsterdam, NL), Marc Cheong (The University of Melbourne, AU), and Mary L. Gray (Microsoft New England R&D Center – Cambridge, US)

The working group on Interdisciplinarity and Computer Science (CS) cultures explored the complexities and nuances of integrating interdisciplinary approaches in computer science research methods and theoretical frameworks. This cluster, involving participants from CS, anthropology and philosophy, with experience conducting mixed methods studies of AI (Asia, Marc, Mary, and Tobias) looked at the tensions, agreements, and common ground that can develop from merging different disciplinary perspectives and methodologies. The discussions highlighted the challenges and opportunities of fostering interdisciplinary collaboration and the need for a shared understanding and language to bridge the gap between computer science and social sciences/humanities.

One of the central themes of the discussion is the challenge of defining and using terms like "good" and "bad" within interdisciplinary contexts when communicating about AI. The participants quickly realized that these terms carry different meanings across disciplines, leading to potential misunderstandings and miscommunications. To address this, they emphasized the importance of specifying what is meant by these terms in different contexts and developing a common basic language for evaluating iterations of AI that do not assume there is a linear or universal path of improving AI for all users, regardless of context. This shared language would help clarify where disciplinary specificity is needed and where interdisciplinary collaboration can be most effective.

The cluster also explored the integration of qualitative and quantitative methods within computer science. Participants discussed the potential for developing and evaluating models using qualitative methods alone and the need for reflexivity from both social sciences/humanities (SSH) and computer science (CS) about the limits and peculiarities of their ways of knowing and forms of evidence. They highlighted the importance of interdisciplinary teaming at specific points in the development pipeline, imagining pairs of experts from technical and qualitative fields working together step-by-step to negotiate approaches that meet shared goals. This collaborative approach would ensure that both qualitative insights and quantitative rigor are incorporated into AI development.

Another key takeaway from the discussion is the role of participatory (re)design, crowdsourcing, and citizen science in interdisciplinary AI development. Participants emphasized the importance of involving diverse stakeholders in the development process and ensuring equitable terms for their participation. They discussed the potential for deliberative development processes that include input from various stakeholders, including those from civil society organizations, industry, and academia. This inclusive approach would help ensure that AI systems are developed with a broader range of perspectives and are more attuned to the needs and values of different communities. The open question was how, exactly, to sustain these multistakeholder codesign efforts, given market pressures and the lack of meaningful connections with diverse community groups of experts available to CS.

The cluster also addressed the challenges of forming interdisciplinary projects without a shared definition of what counts as generative AI–and what will be recognized, professionally, as meaningful contributions to the field. STS and social science-oriented participants agreed that their qualitative methods are often misunderstood or misused on the CS side and that there is an underappreciation of multiple methodologies. They emphasized the need to understand where qualitative analysis should fit in the development and evaluation pipeline and the critical importance of data provenance for meaningful evaluation. The discussions also touched on the philosophical models of reality and knowledge that are useful for thinking about the evaluation process, particularly in the context of foundation models that lack a typical ground truth.

In summary, the thematic cluster on Interdisciplinarity and Computer Science Cultures highlighted the complexities and challenges of integrating interdisciplinary approaches within AI development. The discussions centered the importance of developing a shared language and understanding, incorporating both qualitative and quantitative methods, involving diverse stakeholders in the development process, and addressing the philosophical and methodological challenges of evaluating AI systems.

4.3 Working Group 3: Culture Encodability (25022)

Rida Qadri (Google – San Francisco, US), Kalika Bali (Microsoft Research India – Bangalore, IN), Beth Coleman (University of Toronto, CA), Fernando Diaz (Carnegie Mellon University – Pittsburgh, US), Huma Gupta (MIT – Cambridge, US), Sara Hooker (Cohere For AI – Toronto, CA), Maurice Jones (Concordia University – Montreal, CA), Emanuel Moss (Intel – Santa Clara, US), Maryam Mustafa (LUMS – Lahore, PK), Alice Oh (KAIST – Daejeon, KR), and Moira Weigel (Harvard University – Cambridge, US)

Working Group 3 (Fernando Diaz, Moira Weigel, Huma Gupta, Rida Qadri, Sara Hooker, Maurice Jones, Manny Moss, Kalika Bali, Alice Oh, Maryam Mustafa, and Beth Coleman) took up the challenges of encoding cultural nuances into AI, particularly how to develop technical interventions to preserve the richness of cultures but also cultural protocols that consider whether we ought to/should encode culture. Participants debated whether increasing data volume or enriching data with context was more crucial, and whether solutions lay solely in data or also in model design and evaluation. There was general agreement on the need for richer evaluation methods, non-data-centric interventions like model optimization

changes and interface design, and further research into effective encoding strategies. This cluster further explored examples of the complexities of capturing cultural specificity and the technical approaches that might be used to enhance how models represent cultural expression.

One of the central themes of the discussion was the fundamental question of what aspects of culture can and should be encoded and what governance mechanisms could direct these socially consequential decisions. Participants emphasized the importance of being specific about what cultural elements are being targeted for encoding and the limitations of existing technical processes. They discussed, for example, the challenges of condensing culturally nuanced language into text and the loss of cultural variance in machine translation. The conversation highlighted the need for a deeper understanding of the polysemy and thickness of culture, such as the different structures of languages and the epistemic shifts that occur within them. For example, indigenous languages often have a higher proportion of verbs compared to nouns, which presents unique challenges for encoding.

The cluster also explored the disagreements and agreements around the need for more data versus the need for thicker, more contextually rich data. Some participants argued that more data is necessary to capture the full range of cultural expressions, while others contended that the focus should be on developing thicker development pipelines that incorporate expertise and context. They discussed the limitations of current models, which often operate on crude metrics and may not adequately represent the richness of cultural data. The conversation also touched on the potential for non-data-based interventions, such as designing models that indicate their positionality and highlight absences in the data.

Another key takeaway from the discussion was the importance of cultural protocols in the encoding process. Participants emphasized the need for guidelines on what should and should not be encoded and how to ensure that cultural knowledge is represented accurately and respectfully, wiithout placing the burden of identifying harms on those who might be the most likely targets of them. Particularly of concern was how to think about generating more data without further surveilling data contributors and, on the other hand, the limits of using synthetic or existing datasets that will degrade in accuracy and temporal relevance. The working group discussed, for example, the challenges of creating relational databases that link cultural data to archaeological expertise and the limitations of such approaches. The conversation also highlighted the need for research on whether cultural interventions at different points in the development pipeline are effective and how to design user interfaces that are culturally sensitive.

The cluster also addressed the issue of data absences and the challenges of representing missing cultural information. Participants discussed the potential for structuring data in ways that make absences more visible and the importance of acknowledging the partiality of model outputs. They emphasized the need for models to be transparent about their limitations and the gaps in their data. The conversation also touched on the ethical considerations of data collection and the potential harms of hyper-surveillance and extraction.

In summary, the Working Group 3 on Cultural Encodability highlighted the complexities and challenges of encoding cultural knowledge within AI systems. The discussions centered on the importance of being specific about what cultural elements are being targeted for encoding, developing thicker development pipelines, and adhering to cultural protocols. They also emphasize the need for transparency about data absences and the ethical considerations of data collection.

The seminar's second day sent participants away with some homework, asking them to reflect on what artifacts and projects they would want to specifically take forward as outcomes of the Seminar for building a multidisciplinary research agenda.

Participants spent the third and final day of the seminar working in small groups to identify actionable research directions, fueled by group insights from the Firestarter talks and the Day 2 working group discussions. By the end of our last morning together, seminar participants had identified three specific directions to continue from our seminar: 1) development of an agenda-setting document for research on cultural representation and AI; 2) specific projects aimed at large language models for linguistic diversity; and 3) a clever approach to encoding, dubbed "meta-meta data" evaluation and documentation.

5 Open problems

5.1 Future directions based on participant feedback (25022)

Rida Qadri (Google – San Francisco, US), Asia Biega (MPI-SP – Bochum, DE), Georgina Born (University College London, GB), Fernando Diaz (Carnegie Mellon University – Pittsburgh, US), and Mary L. Gray (Microsoft New England R&D Center – Cambridge, US)

The Dagstuhl Seminar, "Towards a Multidisciplinary Vision for Culturally Inclusive Generative AI," received high praise from participants, who appreciated the interdisciplinary nature of the seminar and the diverse range of fields and disciplines represented. Participants found the seminar to be an unprecedented experience that brought together a broad scope of multidisciplinary research and backgrounds not typically found at computing research venues, fostering rich discussions and collaborations that several noted as a first encounter with that discipline. Many participants noted that the seminar inspired new ideas for their research, development, or teaching.

One of the most frequently mentioned positive aspects of the seminar was the high quality of attendees and the organization of the event. Participants appreciated the format, which included firestarter presentations and ample time for informal conversations over meals and coffee. These informal discussions were seen as highly generative, leading to meaningful exchanges and the development of new ideas. The interdisciplinary and cross-cultural focus of the seminar was also highlighted as a significant strength, with participants noting that it allowed for a deeper understanding of the challenges and opportunities in the field of generative AI and cultural diversity.

However, participants also provided several recommendations for changes to improve future seminars. One of the most common suggestions was to extend the duration of the seminar. Many participants felt that the seminar was too short, given the challenges of developing a shared language for key contested concepts like "cultural representation." They felt an additional day would have allowed for deeper engagement and more thorough exploration of the topics as well as opportunities to establish solid next steps for collaboration. This extension would also provide more time for informal conversations during the day, which participants found to be highly valuable.

Another recommendation was to include a broader range of voices in future seminars. Participants suggested incorporating more representatives from civil society organizations, funders, and philanthropists, as well as increasing the representation of researchers from regions such as Africa, China, and Latin America. Additionally, some participants recommended involving more junior researchers and providing more opportunities for socializing and personal discussions.

Participants also highlighted the importance of including more detailed case studies and examples of interdisciplinary work in future seminars. They felt that this would generate more concrete and detailed ideas about extending this kind of research. Some participants suggested that the seminar could benefit from more explicit links to other participants' work and position statements beforehand, as well as pre-meeting introductions to help participants get to know each other before arriving at the seminar.

The survey results revealed the success of the Dagstuhl Seminar in fostering interdisciplinary collaboration and generating new ideas. Participants appreciated the unique environment provided by Schloss Dagstuhl.

Participants

- Virgilio Almeida
Federal University of Minas Gerais – Belo Horizonte, BR
- Elisabeth André
Universität Augsburg, DE
- Naveen Bagalkot
Manipal Academy of Higher Education – Bangalore, IN
- Kalika Bali
Microsoft Research India – Bangalore, IN
- Asia Biega
MPI-SP – Bochum, DE
- Tobias Blanke
University of Amsterdam, NL
- Georgina Born
University College London, GB
- Anita Say Chan
University of Illinois at Urbana Champaign, US
- Marc Cheong
The University of Melbourne, AU
- Beth Coleman
University of Toronto, CA
- Hal Daumé III
University of Maryland – College Park, US
- Fernando Diaz
Carnegie Mellon University – Pittsburgh, US
- Catherine d'Ignazio
MIT – Cambridge, US
- Giovanna Fontenelle
Wikimedia – Sao Paulo, BR
- Tarleton Gillespie
Microsoft New England R&D Center – Cambridge, US
- Mary L. Gray
Microsoft New England R&D Center – Cambridge, US
- Huma Gupta
MIT – Cambridge, US
- Sara Hooker
Cohere For AI – Toronto, CA
- Maurice Jones
Concordia University – Montreal, CA
- Emanuel Moss
Intel – Santa Clara, US
- Maryam Mustafa
LUMS – Lahore, PK
- Alice Oh
KAIST – Daejeon, KR
- Rida Qadri
Google – San Francisco, US
- Noopur Raval
University of California at Los Angeles, US
- Darci Sprengel
King's College – London, GB
- Molly Steenson
American Swedish Insitute – Minneapolis, US & Carnegie Mellon University – Pittsburgh, US
- Harini Suresh
Brown University – Providence, US
- Moira Weigel
Harvard University – Cambridge, US

Report from Dagstuhl Seminar 25031

Addressing Future Challenges of Telemedicine Applications

Matias Volonte[*1], Andrew T. Duchowski[*2], Nuria Pelechano[*3], Catarina Moreira[*4], and Joaquim Jorge[*5]

1 **Clemson University, US.** mvolont@clemson.edu
2 **Clemson University, US.** duchowski@clemson.edu
3 **Universitat Politècnica de Catalunya, Barcelona, ES.** npelechano@cs.upc.edu
4 **University of Technology Sydney, AU.** Catarina.PintoMoreira@uts.edu.au
5 **INESC-ID Tecnico Lisboa, Lisbon, PT.** joaquim.jorge@tecnico.ulisboa.pt

Abstract

The Dagstuhl Seminar *"Addressing Future Challenges of Telemedicine Applications"* brought together interdisciplinary researchers to chart a forward-looking vision for remote healthcare delivery. With the rapid evolution of telemedicine technologies, driven by global health crises and enabled by advances in extended reality (XR), artificial intelligence (AI), gaze-based interaction, and embodied conversational agents, this seminar explored the critical intersections of innovation, usability, ethics, and equity. Participants engaged in structured discussions on how immersive and intelligent systems can expand access to care, enhance diagnostic accuracy, and foster human-centered experiences in remote contexts. Key themes included building trust in AI, ensuring inclusive design for diverse populations, leveraging eye-tracking and avatars for personalized interaction, and balancing automation with human expertise. The seminar emphasized that addressing technical, cultural, and regulatory challenges is essential to responsibly shaping the future of telemedicine. Through collaborative dialogue, the seminar laid the groundwork for next-generation healthcare technologies that are explainable, adaptive, and empathetic.

Seminar January 12–17, 2025 – https://www.dagstuhl.de/25031
2012 ACM Subject Classification Human-centered computing → User studies; Human-centered computing → User models; Human-centered computing → Usability testing; Human-centered computing → Interaction devices; Human-centered computing → Mixed / augmented reality
Keywords and phrases Telemedicine, eXtended Reality, Eye Tracking, Embodied Conversational Agents & Avatars
Digital Object Identifier 10.4230/DagRep.15.1.50

1 Executive Summary

Matias Volonte (Clemson University, US)
Andrew Duchowski (Clemson Uvniversity, US)
Nuria Pelechano (Universitat Politècnica de Catalunya, Barcelona, ES)
Catarina Moreira (University of Technology Sydney, AU)
Joaquim Jorge (INESC-ID Tecnico Lisboa, Lisbon, PT)

This seminar gathered experts from *XR Technologies*, *Avatars Assistants*, *UX: Gaze Control and Visual Attention*, and *Data Privacy and Security* fields with the objective of identifying the strengths and weaknesses for delivering healthcare assistance remotely, safely, and efficiently. Experts from these different fields described the state-of-the-art of their area of expertise

[*] Editor / Organizer

Addressing Future Challenges of Telemedicine Applications, *Dagstuhl Reports*, Vol. 15, Issue 1, pp. 50–83
Editors: Matias Volonte, Andrew Duchowski, Nuria pelechano, Catarina Moreira, and Joaquim Jorge
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

and addressed future directions that these technologies should focus on for creating the next generation of healthcare telemedicine systems. The outcomes of the proposed seminar will hopefully be highly relevant to researchers in academia and healthcare as well as to the field of Human-Computer Interaction.

A four-day Dagstuhl Seminar was organized to bring together experts from the fields of extended reality (XR) technologies, artificial intelligence (AI), embodied conversational agents, and eye tracking. The seminar followed a structured daily format designed to foster interdisciplinary exchange and collaborative discussion.

Each day commenced with a series of interdisciplinary presentations in which designated experts provided focused overviews on key developments and challenges within their respective domains – namely, XR technologies, AI, and embodied conversational agents. These sessions aimed to establish a common knowledge base and to contextualize ongoing research efforts. Following the morning presentations, participants were divided into interdisciplinary working groups.

Each group engaged in facilitated discussions on predefined topics, chosen to encourage integration across disciplines and to address open research questions relevant to the seminar's overarching themes. In the final session of each day, all participants reconvened in a plenary meeting. During this session, each working group reported the outcomes of their discussions, highlighting key insights, areas of consensus, and proposed directions for future investigation.

This reporting session served to synthesize the day's activities and to promote cross-group dialogue. This daily structure was maintained consistently throughout the seminar to ensure coherence and cumulative progress across the four days.

2 Table of Contents

3 Talks Overviews

The following section provides a detailed overview of the organizers' activities across the duration of the seminar. Each day's responsibilities and tasks are outlined to highlight the structure, coordination, and facilitation efforts undertaken by the organizing team. These summaries offer insight into the planning and execution process that supported participant engagement, interdisciplinary exchange, and the overall success of the seminar.

3.1 Telemedicine

Joaquim Jorge (University of Lisbon, PT, jorgej@tecnico.ulisboa.pt)

The presentation examined how Virtual and Augmented Reality (VR/AR) transformed healthcare and rehabilitation. It highlighted practical applications such as enhancing radiology diagnostics, enabling interactive rehabilitation, and advancing surgical planning and navigation. Challenges were addressed, including high implementation costs, hardware limitations, accessibility issues, and data privacy concerns. Solutions showcased included real-time feedback systems and immersive education tools to improve patient outcomes and professional training. The session also explored how VR/AR was integrated into clinical workflows and examined its convergence with AI to enable precision medicine and personalized care. Telemedicine and remote diagnostics were discussed as transformative areas, demonstrating how these technologies bridged gaps in healthcare delivery. The presentation concluded by emphasizing the need for collaboration, innovation, and rigorous validation to ensure that VR/AR achieves its potential to enhance patient outcomes, engage users and expand accessibility in medical science.

3.2 Intersection of AI with VR

Catarina Moreira (University of Technology Sydney, AU, Catarina.PintoMoreira@uts.edu.au)

The presentation explored how Artificial Intelligence (AI) and Extended Reality (XR) are reshaping healthcare, focusing on radiology as a key domain for these advancements. Highlighting applications like improving diagnostic precision, immersive training for radiologists, and expanding global access to healthcare, it also emphasized the role of explainability and trust as foundational elements for AI adoption in healthcare systems.

A key aspect was the innovative use of knowledge graphs and large language models (LLMs) to achieve explainability in medical imaging. These technologies bridge the gap between complex radiological data and actionable insights, making medical findings more interpretable for both clinicians and patients. Knowledge graphs organize relationships between radiological findings, anatomical structures, and clinical conditions, providing a clear, structured representation of the reasoning behind diagnoses. For example, slides demonstrated how knowledge graphs can connect concepts such as cardiothoracic ratios, pleural effusions, and associated opacities, offering an intuitive, visual explanation of why a heart may appear enlarged on an X-ray.

Large language models, like GPT-4o, were shown to complement knowledge graphs by transforming radiology reports into interactive and understandable interfaces. In projects such as ReXplain, LLMs summarize key findings, link them to annotated images, and use avatar-based interfaces to deliver patient-friendly explanations. For instance, a radiology report describing a pneumothorax or atelectasis is paired with annotated medical images, 3D organ renderings, and video-based explanations. This approach ensures that both medical professionals and patients can better understand diagnostic outcomes, fostering trust in AI-driven systems.

The presentation also emphasized interactive explainable interfaces powered by these technologies. Through visualization tools, radiologists can explore layered explanations, starting with a high-level summary and diving deeper into the relationships between clinical variables. This interactive experience allows for real-time query answering, enhancing confidence in diagnostic conclusions while enabling clinicians to communicate findings effectively to patients.

Another core innovation was the use of digital twins, virtual replicas of organs or systems, generated through AI segmentation models. These models reconstruct 3D visualizations from medical imaging, such as CT scans, providing an immersive and detailed view of patient-specific anatomy. For example, a 3D colon reconstruction created from AI-predicted segmentations was showcased as a tool for enhanced diagnostic and procedural planning.

The session further explored how AI and XR bridge global healthcare disparities by extending radiological expertise to underserved areas. Portable XR systems allow practitioners in remote regions to collaborate with urban radiologists using 3D medical models. This approach empowers non-experts with decision support, fostering global health equity.

In conclusion, the presentation highlighted the transformative potential of AI and XR in healthcare. By integrating innovative technologies like knowledge graphs, LLMs, and interactive explainable interfaces, it demonstrated how these tools enhance diagnostic precision, improve medical training, and build trust in AI-driven healthcare systems. The future of medicine, as depicted, is one of collaboration, transparency, and accessibility, driven by the synergy of cutting-edge technologies.

3.3 Gaze Interaction in XR

Andrew Duchowski (Clemson University, US, duchowski@acm.org)

Andrew Duchowski's presentation on "Gaze Interaction in XR" reviewed gaze-based interaction, distinguishing eye movement analysis from synthesis in virtual reality, games, and other applications. His focus was on five forms of gaze-based interaction: diagnostic (off-line measurement), active (selection, look to shoot), passive (foveated rendering, a.k.a. gaze-contingent displays), expressive (gaze synthesis), and assistive (subtitling). Diagnostic interaction is the bread and butter of serious applications such as training or assessment of expertise. Active interaction is rooted in the desire to use the eyes to point and click, with gaze gestures recently growing in popularity. Passive interaction is the manipulation of scene elements in response to gaze direction, with an example goal of improvement of frame rate. Expressive eye movement centers on synthesis, which involves the development of a procedural (stochastic) model of microsaccadic jitter, embedded within a directed gaze

model, given goal-oriented tasks such as reading. Assistive technologies are used to expand inclusive experiences, ranging from assisting the Deaf and Hard-of-Hearing with subtitles, or the Blind and Visually Impaired with Audio Description driven by gaze scanpaths. In discussing each form of interaction, Duchowski briefly reviewed classic works and recent advancements and highlighteed outstanding research problems.

3.4 Virtual Humans for Telemedicine: Embodied Conversational Agents and Avatars Assistants

Matias Volonte (Clemson University, US, mvolont@clemson.edu)
Nuria Pelechano (Universitat Politecnica de Catalunya (UPC) – Barcelona, ES, npelechano@cs.upc.edu)

This seminar presentation was structured into two distinct yet interconnected topics: Embodied conversational Agents and Avatars for medicine. In the first segment, the discussion focused on the dual role of Embodied Conversational Agents (ECAs) as both barriers and facilitators in the delivery of telemedicine services. Specific attention was given to how attributes such as verbal and non-verbal communication, cultural alignment, and user-agent rapport can either enhance or hinder the accessibility, efficacy, and inclusivity of telemedicine platforms. Case studies and empirical data from clinical trials involving ECAs were highlighted, illustrating their diverse applications in healthcare, from patient education to therapeutic interventions and chronic disease management.

The presentation expanded the scope by exploring cutting-edge technologies that can augment the capabilities of ECAs to create more seamless and intuitive interactions. Eye-tracking technology was examined as a tool to deepen understanding of user behavior and engagement, providing real-time feedback for tailoring agent responses. The potential of extended reality (XR) systems, including virtual and augmented reality, was discussed in the context of creating immersive environments that facilitate trust and enhance patient-provider communication. Additionally, the talk emphasized the importance of robust cybersecurity measures to safeguard sensitive patient data and ensure trust in ECA-mediated interactions. Together, these topics underscored the interdisciplinary challenges and opportunities in designing ECAs for telemedicine, offering a forward-looking perspective on their role in shaping the future of healthcare delivery.

The second part of this talk covered how avatars that represent the user can be used to either perform training or to meet patients remotely. For the case of training, the expert can show step-by-step instructions with the use of animated self-avatar that can accurately follow the user movements. In the case of remote telemedicine, the doctor could visit patience by being embodied in an avatar that can communicate not only with speech but also using nonverbal communication through gesturing. This multimodal communication allows one to build trust with the patient, and thus the presentation also covered several aspects that affects ethics and trust, such as the level of rendering quality needed, or the need to match gender or race with the patient to improve trust.

4 Day 1: Extended reality (XR) for Telemedicine

4.1 Overall description

The discussion explored the effectiveness of extended reality (XR) for training and discussed key considerations for measuring its impact, such as performance metrics and other evaluative factors. The discussion revolved around learning how medical staff in training learned intubation procedures. In the discussion it was examined how to ensure that training in virtual environments translates to real-world performance and highlights the advantages of technology-mediated training. Challenges like latency and communication issues in remote training are also addressed.

4.2 Key Points from Immersive Training Discussion

- **Immersive Education:** Education is framed in terms of two primary goals – the acquisition of knowledge and the development of procedural skills.
- **Main Challenges in Training:** Traditional training methods face several obstacles, including the high cost of equipment and limited availability of resources such as manikins. For instance, observing and practicing procedures like intubation often depends on whether resources are accessible.
- **Challenges with Immersive Technology:** Fidelity in virtual training environments is critical to ensure that procedural skills learned in XR are transferable to real-world scenarios. Intubation was discussed as a key example requiring high fidelity for effective skill transfer.
- **Alternatives to Virtual Training:** Augmented reality (AR) can be used to enhance physical manikins, allowing learners to visualize internal anatomical structures during procedures, thereby improving comprehension and effectiveness.
- **Simulators:** Simulators are valuable tools for procedural training. Laparoscopy was presented as a well-established example where simulation has had measurable educational benefits.
- **Serious Issues:** The talk also addressed risks associated with procedural errors, especially in critical interventions such as intubation, where mistakes can have fatal consequences.
- **Potential Projects:** Future project ideas include developing virtual monitors to assist in diagnostics and real-time information visualization. Such tools could be particularly useful for anesthesiologists and might employ gaze and gesture-based control mechanisms.

4.3 Challenges and Opportunities

The lack of locally available specialists is presented as a significant challenge, which could be mitigated through telemedicine. Additionally, tools such as video or AR-based systems are proposed to facilitate the transfer of procedural knowledge during surgeries.

Once an XR technology is adopted in a certain learning context, there is the possibility of knowledge transfer to other communities and environments. That transfer should take into account all the stakeholders involved: the initial developers, the targeted physicians, the targeted patients and their communities. The success of a solution in a particular context not necessarily guarantees the success in other environments, due to differences in practices,

cultural beliefs, and resources, just to mention some factors. Critical to the success of any technological solution is working with stakeholders to understand needs and co-developing the solution through participatory design or co-design methods.

For telemedicine in particular, stakeholders include doctors/physicians/nurses and other formally trained medical professionals, informally trained care providers such as family members and hospice staff, technologists and the patients themselves. Communicating across differences in training and background knowledge when modulated by gender, racial, and cultural/ethnic differences is challenging. When the doctor-patient dynamic is added to these differences, there is an exacerbation of the communication challenge.

An example from Colombia comes from a training environment developed for pediatricians in a Hospital in Bogota. The system was well received and it is used by both pediatricians and students to learn and reinforce best practices related to 12 different situations while delivering a newborn. Pediatricians want now to start a dialogue with midwives in a rural area of our country to explore possibilities to adapt this system to their own situations delivering babies while at the same time incorporating some of their best accepted practices.

Additionally, it was noted that while the technologist wants to address these challenges, when this technologist belongs to the mainstream socioeconomic demographic in a WEIRD country, the formally trained physician belongs to the same population, and the patients/informally trained caregivers do not, this creates additional communication challenges in the design process, making it harder to identify where the technology can benefit.

The technologist typically wants to improve/address these challenges. When this technologist belongs to the WEIRD population, and the doctor belongs to the WEIRD population and the patients/caregivers do not, this is its own communication challenge in the design process. This can make it harder to identify where the technology can benefit.

An example from Colombia comes from a training environment developed for pediatricians in a Hospital in Bogota. The system was well received and it is used by both pediatricians and students to learn and reinforce best practices related to 12 different situations while delivering a newborn. Pediatricians want now to start a dialogue with midwives in a rural area of our country to explore possibilities to adapt this system to their own situations delivering babies while at the same time incorporating some of their best accepted practices.

4.4 Applications for Telemedicine

In the case of medicine, one scenario is a surgical procedure performed by a local team whose surgeon lacks the required expertise. A remote expert surgeon oversees the procedure through a remote AR interface that allows them to send guidance through virtual annotations. The local team can communicate with the remote through a voice channel and observe the annotations through an AR interface with 3D registration of the operated organ. Meanwhile, a cohort of students passively observes the procedure from the vantage point of the remote expert. Gaze data from the remote expert is rendered to the local team and also to the student group, while gaze data from the local surgeon is relayed to the remote expert.

An asynchronous example could be a virtual space for mental health awareness. Both patients and physicians could use the space at their own convenience. Patients could relate to the actions and situations of other patients. Physicians could get data from the system about the mental state of the patients and find ways through the system to aid them. There could be differences in the information that each participant sees from others, due to privacy issues or due to interests or focus. VR is beneficial in this scenario because it allows people in different regions to be together, thus helping people in isolated areas.

Other practical examples could include:

- Psychology therapy / phobias treatment: This has been widely explored in the past and has been proven to be a very efficient and successfully way of treating psychological disorder, taking advantage of the fine tuning of the scenarios that we can achieve with VR
- Collaboration in Rural Areas: providing access to expert knowledge through VR for areas that are isolated. This could be used for training, for real-time medical advice, and also for remote visits to patients
- VR – mediated collaboration between doctor(s) and patient(s) via XR: several doctors could be working together, alternatively the patient and the doctor could be virtually together in the virtual room. In this way, patients could communicate by gestures (pointing at where it hurts for example) in a more natural way than videoconference alternatives
- Remote Expert Surgery Guidance: Issues include trust between the local surgery team and the remote expert. Many questions raise, such as: is it important to have an AR avatar representation of the remote expert in the local surgery room?
- Rehabilitation with self-avatars: Self-avatars can be used to provide the Sense of Embodiment (SoE), which leads users to experience the virtual body as if it were their own. This can have a very strong impact on the participant, which can be used for rehabilitation to trigger effects such as the virtual mirror to trigger the mirror neurons that improve mobility (by observing the virtual representation capable of correctly moving and arm or leg for which the user has mobility problems). Embodiment can also be employed to trigger the proprioceptive drift of body parts (the feeling that our virtual hand/arm/feet is in a different location), which can be used to overcome pain thresholds by making the patinet believe he/she still has a pain-free range of motion.

4.4.1 Categories / Dimensions

There are many aspects that need to be taken into account when discussing the concept of virtual togetherness. This working group discussed the following taxonomy for the use of avatars in telemedicine to achieve the feeling of copresence (being in the virtual environment with other people). The group proposed the following taxonomy:

- **Synchronous vs. asynchronous**: most situations that require collaboration will need to by synchronous. However there can be situation such as trainning applications, where the expert could have recorded the entire session and then others can re-play it asynchronously. In this case there will not be able to perform Q&A directly with the expert, but it could be helpful for trainning situation that require repetition.
- **Embodied vs. non-embodied**: embodied implies that a real user is driving the avatar in terms of movement, decisions, audio, etc. However we could have training scenarios with an LLM autonomous agent which would not be embodied by a real participant.
- **Virtual environment vs. telepresence augmented reality**: XR involves both situation, fully immerssive VR applications, or else augmented reality situations, where the patient is in the comfort of his/her home, and the virtual doctor is overlayed with the real environment.
- **anonymity vs. non-anonymous**: this refers to whether anonymity matters or is needed. It could have an impact on several aspects of the avatar, such as the appearance, or voice.

- **Therapist-guided vs discussion groups**: If there is a therapist or doctor driving the session or medical visit, or we have a discussion group where several patients share their experience.
- **dyad vs. group**: whether we have only two users, or whether we have a larger group.
- **training vs. part of a procedure**: we could either have an expert explain how to perform a specific surgery or perform the procedure on a patient while other doctors are virtually attending.

4.5 Challenges and barriers

There are many challenges related to this topic. The main issues that were discussed in this working group were:

- What is the **impact of technology** barriers such as latency and registration errors? As with any other technology that requires remote assistance or collaboration, it is necessary to take into consideration the impact that latency due to low bandwidth could have on VR applications. Unfortunately, this is still a problem in many rural areas or third world countries, so when exploring the use of VR applications in telemedicine, we should take this into consideration. The design should recognize that there will always be technical issues and go around it. This is not a new problem, similar issues appear in fields such as aerospace or videogames. Therefore, the first step should be to examine existing solutions in those fields to explore their applicability in telemedicine.
- **Current software limitations**: The lack of standard tools and generalizability of tools for re-purposing, imposes a huge impact when it comes to the quick development of new applications. There is a need for tools that can be reused and easily imported from previous applications.
- **Collaboration / shared virtual spaces**: what are the best solutions for collaboration in VR. For example, the continuum from having multiple VR avatars adhering to social proxemic conventions, over having multiple semi-transparent self-overlapping avatars, to multiple users fully overlapping to the extent of embodying the same avatar (co-embodiment). Many aspects need to be considered when having several avatars sharing the same virtual space, such as proxemics or social rules. When sharing the virtual space, what is convenient of good for the patient may not be the same for the doctor, so it is necessary to find a good balance between the needs of each type of user (e.g. data recording, gaze tracking, body motion traking. etc.)
- **Trust and correct understanding**: In the real world, we know that appearanced and movements have an impact on building trust and achieving a good understanding by having non-verbal communication supporting the audio information. These aspects need to be correctly incorporated in VR when having self-avatars to ensure trust and mutual understanding from the medical expert to the patient.
- **Change established procedures to the ones with new technology**: the medical field has very strict protocols for all procedures that must be followed. Some of them may not be easily adapted to new technology, so alternatives would have to be explored for the purpose of implementing solutions based on VR.
- **Deal with large amounts of data**: new solutions for large data storage would need to be evaluated, taken special care about data protection, privacy, ethics and other legal considerations. It is also necessary to develop robust solutions for deleting persistent data/recording when necessary.

- **How to maintain equipment and systems**: Doctors should not have to learn how to configure systems. VR solutions need to be plug-and-play, meaning very easy to learn and use. There should not be an assumption of any computing competence for the doctors, instead hospitals would need to hire technology experts to rapidly handle any issue with VR. This should not imply a big barrier, given that most hospitals already have complex computer systems that need full time employers to manage and solve problems.
- **research prototypes vs. real applications**: Moving from research prototypes to more complex, flexible yet robust solutions that can work for real applications. Many issues are not considered in prototypes, for example the variety of scenarios and medical cases that need to be taken into account. The final users (i.e. the doctors) should participate in the design process and authoring of scenarios without the need of having advanced computer science skills.
- **dealing with interruptions**: when being in an in-person meeting we are less likely to allow interruptions from our environment or personal life. When in online meetings, this is not the case specially when people are working from home.

4.6 Opportunities

Despite all the barriers and challenges previously mentioned, VR offer a great ammount of opotunities in the field of telemedicine. Some of the main opportunities that were discussed by the group are:

- **Enhanced point of view**: in VR training, we could avoid occlusions of the instructor by seeing only semitransparent hands. This would provide an enhanced learning experience with respect to the real world, by allowing the practicioner to see the procedure from the point of view of the expert, instead of looking over the shoulder or having the point of interest occluded by the experts' hands.
- **Virtual perspective taking**: in the context of a shared virtual space with users having different roles within a procedure (e.g. patient, nurse, doctor, admin, etc.), VR offers a perfect platform for understanding each other perspectives. For example, a surgeon understanding the perspective of the nurse, putting the doctor in the shoes of the patient when receiving difficult news, etc. This aspect of perspective taking has been successfully applied to other application areas, such as understanding racial problems, or domestic violence.
- **Variety of visualizations for each collaborator**: VR offers the oportunity to completely customize the visualization to the preferences, needs of each type of collaborator.
- **Leaning opportunities**: VR offers a platform to learn what happened (if everything is recorded). A challenge that arises here involves privacy and how to ensure that persistent data will protect vulnerable populations, such as patients, etc...

5 Day 2: AI and XR for Telemedicine: Addressing Challenges in Trust, Personalization, and Adoption

The session titled "XAVIER: Explainable AI and Virtual Reality for Enhanced Radiology" set the foundation for the seminar's discussions. This talk explored the transformative potential of combining AI-driven explainability techniques and XR interfaces to enhance radiological workflows and patient outcomes. Key topics included:

- **Knowledge Graphs for Explainability:** Demonstrating how knowledge graphs structure relationships between radiological findings, anatomical structures, and clinical conditions, making diagnostic reasoning interpretable for clinicians.
- **Large Language Models (LLMs) for User Interfaces:** Highlighting how LLMs like GPT-4o can summarize radiology reports, link findings to annotated images, and provide interactive, patient-friendly explanations.
- **Digital Twins and Immersive Visualizations:** Showcasing AI-driven segmentation models that reconstruct 3D anatomical structures for detailed diagnostic and procedural planning.
- **Global Health Equity:** Discussing how portable XR systems and collaborative platforms can extend radiological expertise to underserved regions, empowering non-experts with decision support.

The session aimed to demonstrate how these tools build trust, improve diagnostic precision, and enhance communication between clinicians and patients.

5.1 Breakout Groups and Facilitation

To encourage focused discussions, the participants were divided into breakout groups centered on five pre-identified topics. These topics emerged through interactive polling facilitated by *Wooclap*, where participants voted on priorities and proposed additional themes. LLMs were used to analyze and synthesize participant input, ensuring the breakout topics captured the most pressing challenges and opportunities in telemedicine.

Each breakout group was structured as follows:

- **Topic Briefing:** An initial overview of the topic was provided to establish a common understanding and align goals.
- **Guided Discussion:** Moderators guided discussions using prompts derived from the poll results, encouraging participants to share insights, propose solutions, and debate challenges.
- **Documentation:** Designated note-takers recorded key points and synthesized actionable outcomes.
- **Plenary Reporting:** Each group presented their findings to the broader seminar, sparking cross-group dialogue.

5.2 Relevance of the Breakout Topics

The five breakout topics were carefully chosen to address critical areas where AI and XR could significantly impact telemedicine:

- **Trust, Transparency, and Explainability:** Building trust in AI-driven XR systems is fundamental for adoption. The discussion focused on effective methods for explaining AI outputs, such as provenance tracking and accuracy metrics, to foster confidence among clinicians and patients.
- **Balancing Automation and Human Expertise:** As automation becomes more prevalent, participants debated the boundaries of AI autonomy and the role of clinicians in ensuring patient safety while leveraging AI's capabilities.

- **Personalization and Context Awareness:** Personalized care is a cornerstone of effective telemedicine. This group explored how XR interfaces can adapt to cognitive, physical, and emotional needs, addressing barriers to delivering tailored healthcare at scale.
- **Ethics, Data Protection, and Policy:** Ethical concerns and regulatory compliance are critical to ensuring safe and equitable telemedicine adoption. Discussions focused on frameworks for privacy protection, data ownership, and guidelines for ethical AI use.
- **Usability, Training, and Change Management:** The group tackled practical barriers to adoption, such as usability challenges, training requirements, and organizational resistance, proposing strategies to equip healthcare professionals with the necessary skills.

5.3 Trust, Transparency, and Explainability

This discussion group highlighted the complex interplay between technological design, user perceptions, and ethical considerations in fostering confidence in AI-driven XR tools for telemedicine. The participants examined the dynamics of trust-building, methods for achieving explainability, and the critical role of human oversight in medical decision-making.

5.4 Proposed Methods for Fostering Trust

Participants identified several strategies for cultivating trust among patients and clinicians. *Provenance tracking* emerged as a priority, emphasizing the importance of providing clear, verifiable pathways for generating diagnoses or recommendations. For instance, an AI tool could display the series of inferences leading to its conclusions, akin to an annotated flowchart of reasoning. As one participant noted, this approach helps "reduce uncertainty by revealing the decision-making process." Similarly, accuracy metrics – such as confidence intervals or comparisons with known benchmarks – can offer tangible evidence of reliability.

Another suggestion involved designing interfaces that enable real-time interaction with the AI system. Participants discussed layered explanations, where users can start with a summary and drill down into deeper levels of reasoning as needed. This functionality allows users to engage with AI outputs at a level that matches their expertise, fostering a sense of control and understanding.

5.5 Exploring Broader Scenarios: Trust Beyond AI Systems

To extend the discussion beyond AI-specific tools, participants conducted a thought experiment on trust and transparency in telemedicine more broadly. One scenario considered the integration of 360-degree cameras in virtual consultations. By enabling clinicians to observe a patient's living environment, such systems could provide contextually rich data to support diagnoses. For example, clinicians managing stroke rehabilitation might assess whether a patient's home environment is conducive to recovery, identifying hazards or necessary modifications.

However, participants acknowledged that increased observational capabilities raise critical privacy concerns. The concept of shared spaces was a recurring theme, with one participant remarking, "The patient may not fully understand what is visible to the clinician, and that

asymmetry can undermine trust." The group agreed that transparency about what data is being captured and how it will be used is essential for maintaining both patient autonomy and confidence.

5.6 Challenges Identified: Privacy, Proximity, and Perceptions

The discussion underscored the nuanced relationship between proximity and trust. Remote consultations, while convenient, can sometimes lead to misunderstandings or assumptions that may not arise in face-to-face interactions. For example, a clinician observing a disorganized home environment through a camera might make assumptions about a patient's adherence to treatment without understanding their circumstances fully. Participants highlighted the risk of such misjudgments, stressing the importance of empathetic design that accounts for these social and cultural dynamics.

Privacy was another dominant concern. Participants debated the trade-offs between providing clinicians with more comprehensive data and protecting the patient's right to privacy. One suggestion was to develop systems that allow patients to control what the clinician can observe, providing a sense of agency while still enabling effective care.

5.7 Role of Human Oversight

The concept of a "human in the loop" emerged as a critical point of discussion. Participants agreed that in high-stakes medical contexts, human oversight must remain central to decision-making. However, they also explored how AI systems could enhance this oversight by providing traceable, contextualized explanations. For example, a collaborative AI system might present a suggested diagnosis with a rationale that includes supporting data and references to medical literature, allowing the clinician to make an informed judgment.

Participants debated whether human oversight should always involve the capacity to override AI outputs or whether trust in the system could sometimes allow for autonomous actions. Some argued that enabling clinicians to question and modify AI recommendations is essential for maintaining trust, while others suggested that in certain low-risk scenarios, automation might be more efficient without undermining confidence.

5.8 Insights and Reflections

The group's discussions revealed a shared understanding that trust is not simply about technological robustness but also about transparency, empathy, and respect for user autonomy. By designing systems that are both explainable and interactive, AI-driven XR tools can bridge the gap between complex algorithms and human decision-making. At the same time, addressing challenges such as privacy concerns and asymmetries in information requires ongoing attention to ethical design principles and user-centered approaches.

5.9 Balancing Automation and Human Expertise

This group examined the integration of AI into clinical workflows, emphasizing how automation can enhance human capabilities rather than diminish them. Discussions revolved around delineating the roles of AI in healthcare, addressing cultural and ethical implications, and exploring the limitations and opportunities of automation in a field rooted in trust, empathy, and accountability.

5.10 Defining Roles for AI in Healthcare

Participants categorized the potential roles of AI into three broad categories, reflecting varying degrees of autonomy and complexity:

- AI Task Facilitator: This role involves automating repetitive and routine tasks such as collecting patient data, scheduling appointments, or triaging cases based on urgency. Participants agreed that this role poses minimal ethical concerns and is widely accepted by clinicians. However, ensuring transparency about the data collection process and how the triaging criteria are determined remains critical.
- AI Medical Advisor: In this role, AI systems assist clinicians by synthesizing patient data, generating diagnostic insights, and summarizing relevant medical literature. The group discussed how this role requires a high degree of reliability and traceability, as clinicians depend on these insights to make informed decisions. A participant highlighted, "The AI must not only analyze data but also explain its rationale in a way that aligns with the clinician's thought process."
- AI Decision-Maker: This role envisions AI systems operating autonomously in well-defined, low-risk scenarios, such as prescribing standard treatments for minor conditions. While participants acknowledged the efficiency of such systems, they also raised concerns about over-reliance and the potential erosion of clinicians' expertise in routine medical decision-making.

5.11 Framing and Cultural Implications

The framing of AI's role emerged as a critical factor in shaping user expectations and acceptance. Participants debated the implications of referring to AI systems as "AI doctors" versus "AI assistants." While the term "doctor" conveys a sense of authority and expertise, it also implies accountability and autonomy that many felt was inappropriate for AI. One participant remarked, "*Calling it a doctor implies it can be trusted the same way a human doctor is, which is misleading and could distort the patient-clinician relationship.*"

This discussion highlighted the importance of cultural and societal contexts in defining the roles of AI. In some cultures, patients may prefer clear delineations between human and machine expertise, whereas in others, AI might be more readily trusted if framed as an equal contributor to medical care. The group concluded that consistent and context-aware communication about AI capabilities is vital for managing expectations and building trust.

5.12 Empathy and the Limitations of AI

The group also examined the potential for AI to address issues such as stigma in healthcare. Participants noted that AI systems, by their nature, lack the biases and judgments often associated with human interactions. This neutrality could make AI particularly effective in sensitive areas like mental health, where patients might hesitate to disclose information to a human clinician. However, this advantage is tempered by AI's inability to convey genuine empathy or respond to nuanced emotional cues. A participant emphasized, "*While AI can reduce stigma, it cannot replicate the reassurance of a compassionate human presence, which remains crucial in many clinical contexts.*"

5.13 Regulatory and Ethical Considerations

The group explored the ethical and regulatory implications of increasing automation in healthcare. A recurring concern was the question of liability: if an autonomous AI system prescribes a medication that leads to adverse effects, who should be held accountable – the AI's developers, the healthcare institution, or the clinician overseeing the process? To address this, participants proposed several regulatory measures:

- Confidence Disclosure: AI systems should clearly indicate their confidence levels in specific decisions, enabling clinicians to weigh the AI's recommendations appropriately.
- Traceability: The rationale behind AI decisions should be fully documented, allowing for post hoc analysis in case of errors or disputes.
- Certification Standards: AI systems should undergo rigorous testing and certification processes before being deployed in clinical settings, with standards tailored to the level of autonomy the system is expected to have.

Participants emphasized that while regulatory frameworks can help mitigate risks, they should not stifle innovation. Balancing accountability with adaptability will be key to fostering safe and effective integration of AI in healthcare.

5.14 Broader Insights on Human-AI Collaboration

The discussions highlighted the importance of maintaining the clinician's central role in healthcare decision-making, even as automation becomes more prevalent. Participants expressed concerns about over-reliance on AI, which could lead to a decline in clinical expertise and a loss of critical thinking skills. Drawing an analogy to aviation, one participant suggested, "*Just as pilots are regularly trained to fly without autopilot systems, clinicians should be periodically tested to ensure they can operate without AI support.*"

At the same time, the group acknowledged the potential for AI to act as a safety net, reducing the cognitive burden on clinicians and improving diagnostic accuracy. For instance, AI systems could serve as a second opinion, flagging potential oversights or errors in human judgment. This collaborative dynamic, rather than a hierarchical one, was seen as the most promising path forward.

5.15 Concluding Reflections

The group concluded that the integration of AI into clinical workflows must prioritize complementing human expertise rather than attempting to replace it. By clearly defining roles, addressing cultural and ethical concerns, and implementing robust regulatory measures, AI systems can enhance healthcare delivery while preserving the trust and empathy that are essential to patient care. The discussions underscored the need for ongoing dialogue among stakeholders to navigate the evolving relationship between automation and human expertise in medicine.

6 Day 3: Eye Tracking Technologies

The Dagstuhl Seminar brought together experts to explore the use of eye tracking and gaze-sharing technologies in remote clinical practice, with a focus on improving communication, accessibility, and inclusivity. Discussions highlighted both the potential benefits and the technical and ethical challenges associated with implementing these tools across diverse user populations. The following sections detail the key themes, observations, and recommendations that emerged from these conversations.

6.1 Shared Eye Gaze in Remote Clinical Practice

Eye gaze offers an implicit communication channel that can bridge the physical gap in remote collaboration. In telemedicine, where nonverbal cues are often absent, gaze data can enhance mutual understanding between expert and novice users. One major insight from the seminar was that shared gaze facilitates grounding by confirming that both parties are focused on the same detail. Features like real-time gaze overlays and gaze-based attention indicators were seen as instrumental for improving diagnostic accuracy and collaborative decision-making.

6.2 Perspective Alignment and Virtual Proxemics

Ensuring alignment between the expert's and the local user's viewpoints is critical. This is particularly important in immersive or augmented reality environments, where mismatched perspectives can lead to confusion. Similarly, designers must account for proxemics – the sense of personal space in virtual settings. Poorly managed virtual proximity can lead to discomfort or reduced engagement, especially in high-stakes medical consultations.

6.3 Expert-Novice Communication Gap

Experts may unintentionally omit details they consider obvious, creating challenges for novice users. To address this, the integration of visual overlays, contextual cues, or AI support can help novices interpret the scene more effectively. A feedback loop where gaze cues, speech, and gestures confirm mutual understanding can bridge this expert-novice divide.

6.4 Multimodal Interaction Channels

While gaze is a powerful cue, it becomes even more effective when combined with other modalities such as facial expressions, hand gestures, and speech. The seminar emphasized the importance of designing multimodal interfaces that support naturalistic interactions and mutual understanding in remote healthcare scenarios.

6.5 Challenges and Ethical Considerations

Participants highlighted multiple challenges related to privacy, training, and technical feasibility. Gaze and attention data can be sensitive and reveal underlying health conditions, necessitating strict data security protocols and informed consent. Moreover, real-time systems must function reliably across devices and bandwidth conditions. Clinicians also require training to interpret gaze data accurately, and clear liability structures must be established in the case of miscommunication or diagnostic errors.

6.6 Recommendations for Implementation

The group proposed a set of actionable next steps:

- Develop user interfaces that reflect expert perspectives without overwhelming novices.
- Conduct pilot studies comparing gaze-enhanced sessions with traditional video calls.
- Define ethical frameworks for data use, including consent and anonymization.
- Leverage AI to detect interaction breakdowns or mismatches in shared attention.

6.7 Inclusive and Assistive Eye Tracking for Users at the Edge

The seminar also addressed issues in applying eye tracking to users at the margins – such as neurodivergent individuals, users with cognitive or motor disabilities, and infants. Current systems often fail to account for variability in eye behavior, head shape, pupil distance, or compliance with calibration protocols. Devices may not fit properly, or may not accommodate different gaze patterns and motor behaviors (e.g., in ASD populations).

6.8 Hardware, Software, and Calibration Issues

Poor physical fit, discomfort, and calibration failures were repeatedly identified as barriers to inclusion. Existing eye tracking algorithms and systems are often designed around normative assumptions, failing to generalize across diverse users. The seminar emphasized the need for adaptable calibration protocols, robust hardware, and data-driven models that account for individual variability.

6.9 Data Gaps and the Need for Infrastructure

A major bottleneck for research is the lack of publicly available datasets representing non-normative users. No large-scale repositories currently exist for these populations, and much of the data collected by companies is proprietary. The lack of standardized terminology and absence of de-identification practices further limit the ability to share data responsibly.

6.10 Toward Inclusive Design and Industry Collaboration

To advance the field, participants advocated for:

- Establishing open, crowdsourced databases that capture diverse eye behavior.
- Creating industry standards for inclusive eye tracker design.
- Pushing for business models and incentives that support accessible technologies.
- Encouraging collaborative data collection practices that protect privacy and support reproducibility.

6.11 Unique Challenges in Specific Populations

Special attention was given to young children and users with unique impairments. These groups are often excluded due to head movement, lack of compliance, or outlier biometric features. Customized protocols and adaptable systems will be necessary to ensure that these populations are included in future studies and systems.

7 Day 4: Embodied Conversational Agents and Avatars Assistants

7.1 Introduction

On Day 4, the seminar presentation was structured into two distinct yet interconnected topics, each addressing critical aspects of leveraging embodied conversational agents (ECAs) and Avatars assistants in telemedicine. In the first segment, the discussion focused on the dual role of ECAs as both barriers and facilitators in the delivery of telemedicine services. Specific attention was given to how attributes such as verbal and non-verbal communication, cultural alignment, and user-agent rapport can either enhance or hinder the accessibility, efficacy, and inclusivity of telemedicine platforms. Case studies and empirical data from clinical trials involving ECAs were highlighted, illustrating their diverse applications in healthcare, from patient education to therapeutic interventions and chronic disease management. During this part, eye-tracking technology was examined as a tool to deepen understanding of user behavior and engagement, providing real-time feedback for tailoring agent responses. The potential of extended reality (XR) systems, including virtual and augmented reality, was discussed in the context of creating immersive environments that facilitate trust and enhance patient-provider communication. Additionally, the talk emphasized the importance of robust cybersecurity measures to safeguard sensitive patient data and ensure trust in ECA-mediated interactions.

7.2 Presentation Summary: ECAs and Avatars in Telemedicine

Embodied Conversational Agents (ECAs) are synthetic characters designed to replicate human conversational behaviors. These agents recognize and respond to verbal and nonverbal cues such as gestures, facial expressions, and eye gaze, enabling them to engage in naturalistic interactions. ECAs hold promise in healthcare for delivering personalized, interactive experiences, addressing critical challenges in patient education, mental health support, elderly care, and medication adherence.

Telemedicine witnessed unprecedented growth during the COVID-19 pandemic, with ECAs offering a transformative potential to improve engagement, accessibility, and personalization compared to traditional video conferencing. By leveraging technologies like natural language processing, sentiment analysis, and gaze tracking, ECAs can foster trust, empathy, and social presence, particularly in populations with low healthcare literacy or language barriers.

The presentation highlights key applications of ECAs in telehealth, including their role in therapy for veterans with chronic pain and post-hospital discharge education. It also emphasizes their ability to promote equity in healthcare through culturally sensitive communication and accessibility features, such as voice input for users with low literacy.

Challenges discussed include the "uncanny valley" effect, behavioral imperfections, and the need for cybersecurity measures to protect sensitive patient data. To mitigate these barriers, future advancements in adaptive ECAs are proposed, focusing on multimodal interactions that integrate gaze, body language, and speech to dynamically respond to patient needs.

The presentation concludes with a vision for ECAs to revolutionize telemedicine, overcoming existing barriers and creating more engaging, empathetic, and effective remote healthcare experiences. Future directions include exploring adaptive behaviors, enhancing patient compliance, and improving user experience through innovative technologies.

Avatars were discussed as highly customizable virtual representations, allowing users to modify their appearance or adopt culturally relevant traits. This customization can significantly improve patient comfort and trust, but it also raises ethical concerns regarding identity, representation, and potential misuse. These aspects served as the foundation for the group discussions that followed.

7.3 Group Discussions

After the presentation on the role of Embodied Conversational Agents and Avatar Assistants in telemedicine, the seminar next transitioned into group discussions to explore specific themes in greater depth. Each group focused on a distinct aspect of the topic, fostering diverse perspectives and generating valuable insights. The discussions were structured into four groups, each addressing critical areas: accessibility and equity in telemedicine, trust and ethical considerations in telehealth, technical innovations and challenges in XR technologies, and the use of XR for training and education in healthcare.

Telemedicine presents unique opportunities to bridge gaps in healthcare access, particularly for individuals with disabilities or those in remote areas. Participants in this group highlighted the importance of enabling equitable participation by leveraging avatars and digital agents. For example, avatars that mirror users' real or aspirational appearances can help reduce stigma, allowing users to feel more comfortable and included during medical interactions. Features like text-to-voice systems and subtitles were also identified as critical tools for improving accessibility.

The discussion on Accessibility and Equity in Telemedicine with Avatars explored how avatar representation can promote inclusivity and equity in telemedicine. Key points included:

Representation and Inclusivity: Avatars can allow users to control how they present themselves, addressing needs across genders, disabilities, and cultural backgrounds. However, current avatars lack representation, especially for little people, individuals missing limbs, and gender-neutral options. Attempts at creating gender-neutral avatars often lead to unintended gender perceptions, highlighting challenges in design and cultural biases.

Equity in Interaction: Personalized avatars could help users with disabilities, those facing bandwidth constraints, or individuals with motor or social challenges participate effectively in telemedicine. For example, avatars can bridge gaps in conversational ability or create relatable representations for diverse patients and doctors.

Cultural and Social Impacts: Avatars can challenge stereotypes, such as representing doctors in non-traditional roles (e.g., a child seeing a doctor in a wheelchair). This might also foster social change by addressing biases in patient-provider dynamics.

Personalization in Telemedicine: Stroke rehabilitation agents demonstrated the importance of personalization in conversational style, voice, and memory of patient history. A failure to adapt or remember interactions can harm user trust and satisfaction, emphasizing the need for tailored AI-driven agents.

Continuity for Vulnerable Populations: For patients with memory disorders or high caregiver turnover, digital agents could provide continuity in care. This could be critical in rural areas or communities skeptical of telemedicine due to cultural differences.

Ethical and Emotional Considerations: Avatars that resemble deceased loved ones or culturally insensitive portrayals can cause distress, underlining the need for thoughtful design and deployment.

Insights: Cultural and social diversity must be incorporated into avatar design, ensuring both patients and healthcare providers feel represented. Digital agents capable of maintaining memory and continuity can provide consistent care for patients with chronic conditions, such as Alzheimer's. Personalization in ECAs and avatars should extend beyond appearance to include factors like ethnicity, language, and health literacy, enabling tailored interactions that meet users' unique needs. The group concluded that designing equitable and accessible avatars for telemedicine requires balancing representation, personalization, and cultural sensitivity while addressing technical and ethical challenges.

7.4 Trust, Privacy, and Ethical Considerations in Telehealth

Trust and privacy are foundational elements of telemedicine, especially when sensitive personal and medical data are involved. This group explored how trust can be cultivated through transparency, empathy, and user control over data. Participants emphasized the importance of addressing privacy concerns and ethical dilemmas related to the use of avatars and ECAs in healthcare.

Trust Building. Trust in telemedicine systems is built through empathetic interactions, clear communication about data usage, and involving human oversight in data management. Participants noted that patients often worry about how their data is stored and shared, particularly if it is handled by companies with questionable reputations.

Ethical Concerns. Behavioral data, such as gaze patterns and facial expressions, are valuable for improving telemedicine services but also pose significant risks if misused. For example, poorly anonymized data could lead to breaches of patient privacy or stigmatization. Educating users about what data is collected and how it is used is critical to addressing these concerns.

User's demographics. There is ongoing debate about whether avatars should closely match users' demographic traits, such as age, gender, or culture, to foster trust. While this alignment may boost trust, it also risks reinforcing harmful stereotypes or feeling inauthentic. Participants discussed the balance between personalization and authenticity. Some argued that even if avatars appear artificial, they are acceptable if they improve healthcare outcomes.

Data Protocol procedure. A speaker mentioned that it would be important to develop clear and accessible de-identification protocols for handling sensitive data. Empower patients by providing them with simple tools to manage data sharing and privacy settings.

7.5 Technical Innovations and Challenges in XR for Telemedicine

Participants highlighted the potential of conversational agents to create meaningful interactions by combining high-fidelity facial animations, natural language processing, and adaptive gaze models. Generative models offer a promising avenue for developing realistic avatars that can respond dynamically to user behavior.

Current gaze recognition systems lack robustness in unconstrained environments, making it difficult to maintain realistic interactions. Memory and adaptation capabilities in avatars require significant improvement to enable long-term, personalized interactions. The computational requirements for real-time avatar rendering remain a bottleneck, especially in resource-constrained environments.

Recommendations. Leverage generative AI to improve real-time avatar representation and interaction quality. Conduct longitudinal studies to evaluate the long-term impact of ECAs and avatars on patient engagement, trust, and healthcare outcomes. Explore the use of multimodal inputs – including gaze, body language, and speech – to create more immersive and adaptive telemedicine systems.

Applications. Training healthcare providers in communication and emergency response skills. Providing mental health support and rehabilitation services for patients in remote areas. Offering virtual coaching for managing chronic conditions, tailored to individual patient needs.

The integration of XR technologies, embodied conversational agents, and telemedicine platforms presents several cross-cutting challenges and opportunities that must be addressed to ensure equitable and effective healthcare delivery. Key areas of focus include promoting ethical and cultural sensitivity to represent diverse populations, safeguarding data privacy and security to build user trust, overcoming technological limitations to enable scalability and usability, and adopting user-centered design approaches to create practical and inclusive solutions. These considerations are critical for advancing telemedicine into a future that aligns with the needs and expectations of diverse global communities.

7.6 Summary of Key Perspectives and Discussions

- **Accessibility and Equity:**
  - ECAs and avatars can address barriers faced by individuals with disabilities, low health literacy, or language differences.
  - Features such as text-to-voice systems, subtitles, and culturally tailored interactions enhance inclusivity.
  - Designing inclusive avatars that represent diverse demographics, including gender-neutral and disability-representative options, remains a challenge.
- **Trust and Privacy:**
  - Building trust requires empathetic and transparent interactions, particularly in data handling.

  - Privacy concerns around behavioral data (e.g., gaze patterns, facial expressions) pose ethical challenges.
  - Developing clear de-identification protocols and empowering users with control over data sharing are critical.
- **Ethical Considerations in Representation:**
  - Avatars' customization options improve comfort and trust but raise ethical concerns regarding deceptive representations.
  - Balancing personalization and authenticity is necessary to foster trust without reinforcing harmful stereotypes.
- **Technical Innovations and Barriers:**
  - Advances in conversational agents, high-fidelity animations, and adaptive gaze models enhance telemedicine capabilities.
  - Challenges include computational demands for real-time rendering and improving the robustness of gaze tracking.
  - Long-term personalization and memory capabilities in avatars require further development.
- **Training and Education:**
  - XR environments provide immersive simulations with high-quality anatomical models and haptic feedback.
  - These technologies improve healthcare training and preparedness but demand significant resources and user acceptance.
- **Cybersecurity:**
  - ECAs handle sensitive patient data, necessitating robust cybersecurity measures to mitigate risks like identity spoofing and data interception.
  - Advanced solutions are required to ensure data privacy and maintain user confidence.

7.7 Insights and Forward-Looking Perspectives

- XR technologies, ECAs, and avatars have the potential to enhance engagement, accessibility, and personalization in telemedicine, addressing gaps in traditional healthcare delivery.
- Future research should focus on adaptive ECAs that dynamically respond to user behavior using multimodal inputs like gaze, speech, and body language.
- Longitudinal studies are needed to evaluate the long-term effects of ECAs and avatars on patient trust, compliance, and outcomes.
- Integrating cultural and social diversity into ECA and avatar design is essential for promoting equity and inclusivity.
- Innovations in cybersecurity and de-identification protocols must safeguard sensitive data and foster trust in telemedicine systems.
- XR environments for healthcare training and rehabilitation should be further explored to revolutionize skill acquisition and therapeutic interventions.
- By addressing these challenges, XR technologies and ECAs can drive equitable healthcare delivery for diverse populations.

7.8 Avatars: Diversity in Representing Remote Users

Members emphasized that being represented by an avatar in remote environments provides an opportunity to be fully included and treated as equals, based on the design and capabilities of the avatar that links them to that space. The avatars could even augment their communicative abilities, possibly overcoming limitations, for instance caused by disability. However, there is a need to offer flexibility and choice in avatar representations – some people may prefer a realistic likeness, while others might benefit from different or less stereotypical representations. Concerns were raised about current avatar designs and choices, which often lack diversity (e.g. limited options for avatars that show disabilities or non-typical body types like little people or amputees). It should not be assumed that people with disabilities wish to hide those traits, as that would be a form of ableism. There is also a lack of gender-neutral avatars, which often default to feminine traits. Others noted how avatars could help people express themselves, especially for those from marginalized communities, like transgender individuals, who may use avatars before coming out. Avatars should not only represent patients but also healthcare providers, who are often portrayed in ways that may be intimidating (e.g., as middle-aged, male, healthy individuals). It was suggested that allowing doctors to use avatars that reflect diverse identities could be an intervention for social change, though it's complex and could have unintended consequences.

7.9 Virtual Agents or Characters: Continuity in Patient Care

A member shared his work on developing virtual agents for stroke rehabilitation and stressed the importance of personalization. Patients – especially seniors – want agents with a personality that motivates them. They also expect the agents to remember past interactions, which could reduce frustration for patients. Concerns were also raised about continuity of care for patients with memory disorders (e.g., Alzheimer's), where caregivers' faces change frequently. Virtual agents could provide consistent, comforting interactions, especially in rural areas or for underserved populations. This could lead to the potential for "tele-kits" with virtual agents that could provide continuity of care, even if a patient is admitted to the hospital. This would allow personalized virtual agents to travel with the patient, helping to ensure consistency.

7.10 Hybrid Avatar-Agent: A Unified Face for Remote Service

A unique opportunity presents itself when Virtual Agents could seamlessly become avatars for remote healthcare professionals. As in the previous example of a consistent personalized virtual agent for patients with memory disorders, that same virtual body could become an avatar for a remote physician or other healthcare professional who wishes to interact with the patient remotely. That way, the patient gets remote care delivered through a familiar face, even if the remote professional is not always the same person. Indeed, Entire teams could be represented as the same persona entire teams could be represented as the same persona to patients at home, or even at hospitals. It was also discussed how such an avatar-agent hybrid could gently introduce new people to the patient, for instance before they arrive at the patient's door at home.

7.11 Hybrid Avatar-Agent: Mixed Expertise and Emotional Support

Hybrid avatar-agents could also respond autonomously to patients to address immediate needs that can easily be handled by AI, but would have the capability to bring in experts as needed – all through the unified visual representation. Such hybrid avatar-agents could also maintain positive and emotionally supportive visual behavior towards the patient, regardless of the current emotion or energy level of the professional providing the expertise. Furthermore, the hybrid avatar-agents could even help with or fully autonomously deal with difficult emotional patient interactions, shielding the professionals from some of the "emotional labor" that often places unnecessary and unwanted strain on the professionals.

7.12 Tele-Robotic Avatars: Extending the Work Force

A member shared his experience in Japan with tele-robotics, where remote workers with disabilities control robots that interact with customers at a cafe. Some of them waited on the tables, while others joined guests at the tables to provide social company. This provides unique opportunities for people who otherwise might become isolated at home to integrate with society as valuable members. The particular robotic representation chosen in Japan (e.g. a penguin), or even the idea of such a tele-robotics cafe, may not necessarily be something other cultures might accept. We may need to look into what is possible and culturally acceptable in different countries. The idea was raised that such tele-robotic avatars could be brought into the medical field where they could allow disabled medical professionals to enter hospitals or other places where they could interact with both patients and medical staff, providing consultation or even just emotional or social support.

7.13 Ethical and Practical Considerations for AI in Healthcare

In the working group session on "Balancing Automation and Human Expertise", participants explored the roles AI could take in medical advising, the metaphors that match the respective role, the challenge in balancing out AI and human competencies and the opportunities of using AI in therapeutic and medical contexts.

While the fields of application for AI in medicine and therapy may be wide spread – from its usage in medical institution management over advising medical staff to direct patient contact as an AI companion, advocate or "doctor", the role of AI currently lacks a systematic classification. The working group proposed three potential roles of AI in future telehealth systems along the dimension of responsibility, tightly knit to system complexity: AI as a task facilitator, AI as a medical adviser, and AI as a medical decision maker. AI as a task facilitator might include using it to perform motivational interviews in psychotherapy, to gather information about patients and their background, or to create reports. These tasks demand a relatively low complexity, leaving the main expertise and responsibility with the human operator. The working group expected this level of AI usage as a realistic first step regarding user trust and acceptance. Using AI as a medical adviser increases its complexity, its involvement and also the responsibility gathered towards it. As such, the AI system might summarize information, suggest approaches based on conditions and to proactively inquire with the human expert to ensure they retrace its rationales and agrees with them. One challenge within that role would be implementing the detection of false information or lies by patients instead of assuming truthfulness in all statements. As a solution, the group suggested following existing strict decision diagrams for standard routines.

7.14 Bias in Diagnoses and Advice: Ensuring AI Operational Neutrality

There is a growing concern about bias in AI-driven medical diagnoses and prescription advice. For example, if an AI model is perceived as being developed or owned by a pharmaceutical company, such as a "Pfitzer AI Doctor", trust in its prescription recommendations may be compromised. To address this, AI systems must provide traceable rationales for specific ingredient choices or treatment paths. Furthermore, rigorous certification procedures must be implemented and continuously maintained to ensure impartiality, especially as AI capabilities rapidly evolve in an arms race of technological development.

7.15 Responsibility and Power: Monitoring Between AI and Human Experts

A critical question arises around responsibility: should AI systems monitor human experts, or should human experts monitor AI systems? The answer has deep implications for accountability and legal liability. If an AI advises a course of action that leads to harm, determining who is legally responsible becomes complex. Clear governance structures are required to define authority, establish protocols for dispute resolution, and manage the legal consequences of AI-assisted medical decisions.

7.16 Declining Human Expertise: Mandatory Autonomy Testing for Healthcare Professionals

As AI tools become more integrated into medical practice, there is a risk that human expertise may atrophy. To counteract this, a "Healthcare Autonomy Test" may need to be developed, akin to flight simulator tests in aviation. These tests would ensure that medical professionals retain the ability to operate independently of AI systems, maintaining core diagnostic and decision-making skills even in high-tech environments.

7.17 Changes in Acceptance After Incidents: Design and Communication Considerations

Public acceptance of AI in healthcare can change significantly following incidents or adverse events. These shifts – akin to changes in the Overton window – must be anticipated in the design and marketing of AI systems. It is essential to document moments of public impact meticulously and to communicate AI capabilities in clear, metaphorically appropriate ways. This includes avoiding misleading analogies or exaggerated claims in commercials and ensuring users understand the AI's true role and limitations.

7.18 Matching Hardware Capability to Therapeutical Requirements

The effectiveness of AI in healthcare is also constrained by the quality of supporting hardware. In domains such as radiology or surgical robotics, improved hardware can enhance the AI's analytical, interpretive, reporting, and decision-making capabilities. However, in other

fields like psychology, where interventions are more conversational or cognitive, the demand for high-end hardware may be less critical. Identifying the appropriate level of hardware sophistication for each specialty is necessary for optimizing clinical integration.

8 Day 1-4: Data Privacy and Security

Data privacy and security emerged as a recurring theme throughout the seminar, woven into discussions across all four days. Given the sensitive nature of healthcare data and the increasing reliance on connected, intelligent systems in telemedicine, participants consistently emphasized the importance of ethical data handling, robust security protocols, and user trust. Whether examining gaze-based interfaces, AI-driven diagnostics, or avatar-mediated communication, conversations regularly circled back to the need for secure, transparent, and equitable data practices.

Table 1 Summary of Trust, Privacy, and Ethical Considerations in Telehealth.

Category	Details
Trust, Privacy, and Ethical Considerations in Telehealth	▬ Self-avatars: high fidelity/cartoonish ▬ Deceptive appearance (age/gender/...) good or bad? ▬ Gaze (synthetic)
Trust in Telehealth	▬ Trust builds through empathy, transparent data use, and involving humans in data management. ▬ People worry about privacy – how data is stored, shared, or even mishandled by companies with questionable reputations. ▬ A big question: how do we balance user control over data with the need for enough data to provide accurate care?
Privacy and Ethical Concerns	▬ Behavioral data (like gaze tracking) and health records raise risks, especially if they're misused or poorly anonymized. ▬ Not everyone understands what data collection means, and lower health literacy could worsen this. ▬ Socioeconomic factors play a role: affluent users might demand more privacy, while others don't have the same options.
Avatars in Telehealth	▬ Should avatars match a user's age, gender, or culture? It might boost trust, but there's a risk of reinforcing stereotypes. ▬ Some argue fake personalization could feel inauthentic, while others say it's fine if it improves outcomes. ▬ Participatory design – creating avatars with community input – can help avoid ethical pitfalls.
Emerging Data Types	▬ Gaze, behavior, and other new data types offer insights but come with risks of misuse or over-collection. ▬ De-identification techniques could help, but protocols for how this data is handled are still unclear.

8.1 Privacy in Healthcare Data

The discussion covered the multifaceted issues surrounding privacy in healthcare, AI, XR, ECA, data collection through wearable sensors, applications, usage and regulatory processes. A key focus was the intricacies of data collection in healthcare, particularly how personally identifiable information (PII), such as demographic profile of the user combined with interaction history and other information from doctor summaries, is handled in a secure way to ensure users' privacy and minimize potential risks. The discussions underscored the ethical need for strict protocols and regulations to safeguard sensitive data and protect users. Topics such as the use of biosensors, audio, and video recordings revealed both the potential benefits for health outcomes along with the privacy risks inherent in modern healthcare technology. De-identifying data to minimize privacy risks was emphasized, particularly while exploring behavioral data and ignoring potential implication and the consequences of its misuse. The sources of data as well as its representation when developing content and interaction in telemedicine contexts – if not done in a critical way– can further exacerbate stigma and mental health challenges.

8.2 Trust, Ownership, and Patient Autonomy

Ensuring trust in the healthcare delivery through virtual agents remains as an open challenge to be addressed, along with safe processes to handle the medical records in a telemedicine context. While we acknowledge that patients should have freedom of choice, there is a complexity and added burden inherent to control choices. Surely patients must have the ability to opt in or out of data usage processes, seeking to respect their autonomy. Yet there are aspects of privacy literacy to be considered too. Concerns about power dynamics were raised, particularly in telehealth contexts where patients may feel coerced into sharing data with de facto platforms (ECA, avatars) or organizations without room for negotiation (like health insurance companies). Other issues raised involved the bias on how information is perceive depending on the characteristics of the embodied agents (demographic profile, voice, speech, accent). Companies were urged to prioritize privacy literacy and offer clear terms and conditions, or other applicable measures that are intuitive and accessible to the patient, empowering patients to make informed decisions about their data in a flexible, negotiable way. Building trust with users requires transparency, strong reputations, and good company values. Critical reflection is necessarily to go beyond retroactive strategies that

8.3 Risks and Consequences of Data Usage

The discussion raised also the risks associated with the PII and sharing of healthcare data, opening risks such as misuse for marketing purposes, health insurance premium changes, denial of care, or biased decision-making by health practitioners, insurance companies, or other stakeholders involved. Participants reflected on the growing role and threats involved with the usage of deceptive appearances (e.g. through deep fakes), such as changeable avatars or personalized agents, which can manipulate user trust and introduce cultural biases. Discussions also touched on the burden of privacy risks, especially for marginalized communities or those suffering from mental health conditions, where loss of privacy can deepen social stigma and worsen health outcomes. The cost of free access to healthcare platforms was a recurring theme, as it often comes at the expense of personal data and it may reinforce inequalities in treatment.

8.4 Organizational Responsibilities and Regulatory Needs

Even though nowadays organizations handling healthcare data are required to adopt strict regulations and clearly defined purposes for data usage, privacy risks in a telemedicine contexts are unknown and to minimize errors in handling sensitive information more efforts are needed [1]. Critical reflection on the implications of data sharing and aggregation highlighted the need for a balanced dynamic between organizational interests and user privacy. The role of health insurance companies and their influence on reimbursement and data usage policies has to be scrutinized to prevent harmful processes. Attendees stressed the importance of building a robust regulatory framework to address evolving privacy risks in the context of modern telemedicine operations and maintain trust in the doctor-patient relationship.

8.5 The Role of Personalization and Trust in Healthcare

Finally, the discussions addressed the vast potential of personalization in healthcare, such as tailoring personas in an embodied agent or avatar, and consideration about the intersectionality of individual patients to ensure that the delivery of healthcare meets individual needs without relying on stereotypical and reducionist approaches that not only limit the patient to a single attribute (like nationality) but also exacerbate bias in treatment. While this personalization approach offers promising benefits for enhancing health outcomes, it also poses challenges related to power dynamics and bias. For example patients are more likely to trust an ECA with certain traits, but adopting "preferred" traits will reinforce bias and prejudices even further. Overall, the trust in the medical system remains a cornerstone of patient confidence, but doing so is achievable through community-based approaches relying on a deep understanding of users, as well as an energetic, transparent communication. Organizations must integrate seamlessly innovation with ethical considerations of fairness and trustworthiness, ensuring that personalization does not lead to manipulative practices in the field or prejudice. By fostering trust and prioritizing patient choice, healthcare systems can achieve both technological advancement and equitable care that meets the needs of marginalized populations.

References

1 Vivian Genaro Motti and Shlomo Berkovsky. Healthcare privacy. In *Modern Socio-Technical Perspectives on Privacy*, pages 203–231. Springer International Publishing Cham, 2022.

9 Discussion

The Dagstuhl Seminar *"Addressing Future Challenges of Telemedicine Applications"* served as a multidisciplinary forum that brought together researchers and practitioners to reflect on the current and future challenges of telemedicine technologies. The central themes discussed throughout the seminar revolved around the integration of XR technologies, AI, eye tracking, and embodied conversational agents, with a strong emphasis on accessibility, trust, inclusivity, and clinical relevance.

A recurring thread across all sessions was the necessity for building trust and transparency in AI-driven telemedicine systems. This includes the importance of explainable interfaces that help clinicians and patients understand the reasoning behind AI outputs. Provenance

tracking, confidence scores, and layered explanations were discussed as viable strategies to increase interpretability and foster confidence. These methods aim not only to justify clinical decisions but also to support accountability and ensure ethical standards in remote healthcare delivery.

Another key area of discussion was the role of immersive and extended reality in transforming medical training and remote collaboration. Through virtual environments and avatar-based interactions, clinicians can train or deliver care in shared virtual spaces. These systems provide new ways to represent co-presence and embodiment in therapeutic and educational contexts. However, participants emphasized that such tools must be critically examined for accessibility, particularly among users with cognitive, motor, or perceptual differences. Topics such as gaze calibration issues, hardware fit, and lack of representative datasets were identified as technical and infrastructural challenges that continue to marginalize certain populations.

Ethical dimensions also featured prominently in the seminar discussions. From the need to clearly define the role of AI in healthcare workflows to managing data privacy concerns and regulatory standards, the participants explored the balance between innovation and responsible design. Cultural framing and communication around AI capabilities – particularly in how we label and market AI systems – was also seen as having a direct effect on user expectations and trust.

Participants recognized that current research prototypes often fall short of real-world deployment, particularly in underserved regions where infrastructure may not support high-bandwidth XR applications. Emphasis was placed on ensuring usability, creating inclusive design pipelines, and adopting participatory design methods that empower patients and practitioners alike.

In sum, the discussions illuminated a shared vision: future telemedicine tools must not only be technologically advanced but also ethically grounded, inclusive by design, and tailored to the diverse needs of patients, clinicians, and care ecosystems.

10 Conclusion

The Dagstuhl Seminar *"Addressing Future Challenges of Telemedicine Applications"* successfully convened a diverse community of experts to address the interdisciplinary challenges at the intersection of telemedicine, extended reality, artificial intelligence, and embodied interaction. The four-day seminar fostered meaningful discussions that revealed both the tremendous promise and the complex hurdles that lie ahead in the evolution of remote healthcare systems.

Key takeaways from the seminar underscore the need for solutions that are not only technologically robust but also ethically sound and accessible to all users. Trust-building mechanisms such as explainable AI and transparency in data usage were identified as essential features for widespread adoption. Additionally, participants emphasized the importance of designing for inclusivity – from hardware and software that accommodates non-normative users, to interface designs that support multimodal and culturally sensitive interactions.

Future directions call for collaborative efforts across disciplines. The challenges discussed – from enabling shared gaze in remote diagnostics, to developing empathetic and personalized virtual agents, to establishing regulatory frameworks for AI in medicine – cannot be addressed in isolation. Researchers, developers, clinicians, and policy-makers must work together to ensure that telemedicine technologies evolve with a balance of innovation, responsibility, and empathy.

Ultimately, the seminar provided not just a space for dialogue, but a foundation for ongoing collaboration and actionable research. It laid the groundwork for future initiatives that aim to advance telemedicine as a core pillar of equitable and patient-centered healthcare in the digital age.

Participants

- Gerd Bruder
University of Central Florida – Orlando, US
- Andreas Bulling
Universität Stuttgart, DE
- Joana Campos
NESC-ID – Porto Salvo, PT & Instituto Superior Técnico – Lisbon, PT
- Carolina Cruz-Neira
University of Central Florida – Orlando, US
- Nina Döllinger
Universität Würzburg, DE
- Andrew Duchowski
Clemson University, US
- Pablo Figueroa
Universidad de los Andes – Bogotá, CO
- Justyna Garnier
University of Social Sciences & Humanities – Warsaw, PL
- Vivian Genaro Motti
George Mason University – Fairfax, US
- John Paulin Hansen
Technical University of Denmark – Lyngby, DK
- Nina Hubig
Clemson University, US
- Victoria Interrante
University of Minnesota – Minneapolis, US
- Eakta Jain
University of Florida – Gainesville, US
- Joaquim A. Jorge
University of Lisbon, PT
- Regis Kopper
University of North Carolina – Greensboro, US
- Joseph J. LaViola
University of Central Florida – Orlando, US
- Benjamin C. Lok
University of Florida – Gainesville, US
- Päivi Majaranta
Tampere University, FI
- Belen Masia
University of Zaragoza, ES
- Catarina Moreira
University of Technology – Sydney, AU
- Luciana Nedel
Federal University of Rio Grande do Sul, BR
- Joao Ricardo Nickenig Vissoci
Duke University – Durham, US
- Tabitha C. Peck
Davidson College, US
- Florian Pécune
University of Bordeaux, FR
- Catherine Pelachaud
Sorbonne University – Paris, FR
- Nuria Pelechano
UPC Barcelona Tech, ES
- Daniel Perez-Marcos
MindMaze – Lausanne, CH
- Voicu Popescu
Purdue University – West Lafayette, US
- Nelson Silva
IT:U Interdisciplinary Transformation University – Linz, AT
- Richard Skarbez
La Trobe University – Bundoora, AU
- Jeanine Stefanucci
University of Utah, US
- Hannes Högni Vilhjálmsson
Reykjavik University, IS
- Matias Volonte
Clemson University, US
- Sebastian von Mammen
Universität Würzburg, DE
- Gregory F. Welch
University of Central Florida – Orlando, US
- Gabriel Zachmann
Universität Bremen, DE
- Katja Zibrek
INRIA – Rennes, FR

Report from Dagstuhl Seminar 25032

Task and Situation-Aware Evaluation of Speech and Speech Synthesis

Jens Edlund[*1], Sébastien Le Maguer[*2], Christina Tånnander[*3], Petra Wagner[*4], and Fritz Michael Seebauer[†5]

1 **KTH Royal Institute of Technology – Stockholm, SE.** edlund@speech.kth.se
2 **University of Helsinki, FI.** sebastien.lemaguer@helsinki.fi
3 **Swedish Agency for Accessible Media – Malmö, SE.** christina.tannander@mtm.se
4 **Universität Bielefeld, DE.** petra.wagner@uni-bielefeld.de
5 **Universität Bielefeld, DE.** fritz.seebauer@uni-bielefeld.de

Abstract

Speech synthesis has now reached such human-likeness that its evaluation as a separate entity is no longer meaningful. In this Dagstuhl Seminar, we approach speech and speech synthesis evaluation from a multidisciplinary perspective. Our goal has been to establish a core network to reach all impacted research communities and to provide fundamental directions to develop the new standards of speech and speech synthesis evaluation.

Seminar January 12–15, 2025 – https://www.dagstuhl.de/25032
2012 ACM Subject Classification Human-centered computing → HCI design and evaluation methods; Computing methodologies → Natural language processing; Human-centered computing → Accessibility design and evaluation methods
Keywords and phrases evaluation, human-in-the-loop, speech technology, speech-to-text synthesis
Digital Object Identifier 10.4230/DagRep.15.1.84

1 Executive Summary

Jens Edlund
Sébastien Le Maguer
Christina Tånnander
Petra Wagner

This report documents the program and the outcomes of Dagstuhl Seminar "Task and Situation-Aware Evaluation of Speech and Speech Synthesis" (25032).

The recent advances in deep neural netowrks have pushed the boundaries for synthetic speech to the point where synthetic speech is, in some contexts, indistinguishable from human speech. Alongside a slew of well-known issues with deep fakes, this development raises fundamental questions concerning the evaluation of synthetic speech and its relation to the evaluation of human speech. Human speech and synthetic speech have traditionally been evaluated in different ways, with human speech often serving as an implicit or explicit gold standard for synthetic speech. At the same time, the technical distinction between synthetic and human speech is getting increasingly blurred: human speech is delivered through encoding/decoding processes that changes the signal fundamentally – most notably

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

in applications such as speech-to-speech translation and anonymisation – and voice cloning of recorded speech passes as speech synthesis. We hold that the fundamental question when evaluating speech these days is no longer "How similar to human performance is this?", but rather: "Is this "good" speech?"

The issue is made more complex still by the fact that what constitutes good speech, synthetic or human, is not a trivial question. Finally, standard evaluation methodologies fail to take into account the interdiciplinary nature of speech science and and speech technology through its assumption that one single evaluation metric should satisfy all requirements.

The goal of the Dagstuhl Seminar "Task and situation-aware evaluation of speech and speech synthesis" (25032) was to initiate shift in the different communities impacted by these evolutions. To do so, we gathered a total of 22 renowned researchers from various disciplines (among others: engineering, phonetics, user interface, computer science, speech pathology) to exchange about this fundamental issue. Exchange between groups was encouraged by organisating the seminar around working groups designed to explore the breadth of research fields and applications with stakes in speech and synthetic speech evaluation. This was also intended to encourage more active involvement of the participants both during and after the seminar. This hands-on approach came at the expense of formal talks and panel discussions, which were limited to two talks, both intended to get background "out of the way" and to allow the rest of the discussions to focus on the future.

In the present executive summary, our goal is twofold. Firstly, the presentation of the immediate outcomes of the dicussions that took place during the seminar. In addition, we are convinced that the manner in which the seminar was organized represents an contribution as it presents a process towards fruitful trans- and interdisciplinary exchange leading to a long term impact.

Each day of the three-day seminar was given a broad goal: The first day focussed on background, the second on innovation and solutions, and the third on consolidation and structuring. Each day was further divided into sessions, each with a specific result in mind.

Following a general introduction to the seminar, including Dagstuhl practicalities, the first day started out with three-minute participant presentations. As most of our participants have a long and broad set of experiences in the speech field, they were encouraged to focus their presentations on matters of direct relevance to the seminar, and they were also asked to specify their own interests in relation to the seminar description. Since this seminar was designed to gather and reconcile as much as possible of the collective experiences in speech and speech synthesis evaluation, with few presentations and much discussion and collaborative work, we include all personal statements in this report 4.1. They constitute a fair representation of the morning session. After a following discusssion and a brief introduction to the afternoon sessions, the morning session was concluded.

The afternoon session continued and concluded what we view as the background work in this seminar. First, two talks presented the limits of the current state-of-the-art in speech synthesis evaluation methodologies. A longer session dedicated to an extensive exchange about these methodologies and their limits followed the talks. Although these shortcomings were are well-known within the group of participants, experience has shown that discussion on improvements and evaluation innovation easily get stuck in repeated discussions about the shortcomings of current methods. Our goal was to get these discussions out of the way on day one, and then explicitly avoid going back in days two and three, to ensure that the remaining time of the seminar was dedicated to the exploration of new horizons.

At the end of day one, the organisers presented a set of speech and speech technology areas with suggested use cases, together with group assignments for all participants, in order to allow the participants to muse over these in preparation for the second day.

The second day was dedicated to discover and explore new directions for speech and speech synthesis evaluation. This day was structured around four working group sessions interspersed with flash presentation plenary sessions. The goal of these plenary sessions was not only to fuel the discussion of the following sessions, but also to inform all the participants of the reflexion of each working group. The first three sessions were centred around high-level use cases known to be impacted by the recent evolution in speech synthesis.

The organizers defined the groups by taking into account the background and the interests of the participants as communicated in their personal statements. This strategy proved to be effective as no participants requested to change group. While the first two sessions aimed at developing the initial use cases and what falls under each use case, the third session focused on exploring potential methodologies. Note that the use cases, here, served primarily as a focal point for discussion, designed to capture specific TTS and speech characteristics and requirements as well as cover specific types of applications. Thus the goal was not to create fully fledged and ready-to-use protocols, but to explore what constraints and requirements future methodologies will have to meet.

Informed by these meetings, the organizers then defined a set of five methodology umbrellas which emerged from different use cases, and the last session was dedicated to exploring these umbrellas. For this session, the participants were left free to select the group to join. While this led to an uneven distribution, this also provided space for the participants to focus on some of their more immediate interests.

The last session of the day was a plenary general discussion of the day's events, aiming to allow participants to bring up points of criticism and complementary information.

The last day was dedicated to clean, collect, summarize the information produced previous day, with a clear aim at future work and practical ways to continue the work discussed in the seminar. The result was a less intensive and the organisers also decided to give more freedom for the participants to determine concrete activities they wanted to participate. Working group formed spontaneously to establish and engage in short term solutions to address the current flaws in speech evaluation.

The immediate outcome of the seminar is the establishment of a core network of renowned researchers from various disciplines dedicated to speech and speech synthesis evaluation. Due to its multidisciplinary and breadth, this community can reach a range of different research communities.

A major achievement in the wake of the seminar is a set of new guidelines for reviewing TTS papers with respect to evaluation. Members of the network are part of the organisation committees of Interspeech – the reference conference about speech science and speech technology – and the Speech Synthesis Workshop (SSW), and have promoted – for Interspeech – or enforced – for SSW the use of these guidelines. In addition, a discussion on edits and amendments to the ITU reports on TTS is currently undertaken with ITU.

Several papers are also underway as direct results of the seminar, including a position paper to SSW about the new directions in speech synthesis evaluation and a more substantial survey article to the journal Computer, Speech & Language (CSL).

To increase awareness, we are currently exploring other ways of dissemination such as designing tutorials to be presented at Interspeech 2026, as well as a repeating work shop series. We will also propose a special issue of the journal CSL dedicated to speech and speech synthesis evaluation.

In the longer term, the goal reaches far beyond documenting the current state of speech and speech synthesis evaluation, towards a dynamic process that avoids renewed fossilisation. We believe we have achieved this on three fronts. First, the seminar ensured the transmission

of information between generation of researchers. This is necessary to keep the field cohesive. Second, the seminar brought researchers both from academia and industry. This is critical to ensure that balanced is maintain between the different interests. Finally, the seminar provided the space to not only get new directions but also to have a core set of renowned researchers whose duty is now to impulse this change in their respective communities armed with the different resources provided by the activities of the seminar.

2 Table of Contents

3 Overview of Talks

3.1 Day 1, session 1b. MOS and its limitations and biases

Erica Cooper (NICT – Kyoto, JP)

Main reference Erica Cooper, Junichi Yamagishi: "Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech", in Proc. of the 24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023, pp. 1104–1108, ISCA, 2023.
URL https://doi.org/10.21437/INTERSPEECH.2023-1076

This talk introduced the Mean Opinion Score (MOS) listening test protocol for evaluating synthesized speech, as well as some of the limitations and biases of this protocol. We described the collection of a large-scale MOS dataset covering synthesis systems from 2008-2022 and what we learned about listener preferences and the evolution of speech synthesis technology from this dataset. We presented the VoiceMOS Challenge, a shared task challenge for automatic evaluation of synthesized and processed speech, along with the lessons we learned from three years of running the challenge about which approaches work well for automatic prediction as well as what makes the task difficult. Lastly, we presented a case study on range-equalizing bias, which is the tendency of listeners to use the entire range of the rating scale regardless of the absolute quality of the samples presented to them, and demonstrated that MOS ratings for the same system can vary by over one point depending on the relative quality of the other samples included in the test.

3.2 Day 1, session 1b. The limits of the Mean Opinion Score for speech synthesis evaluation

Sébastien Le Maguer (University of Helsinki, FI)

Joint work of Sébastien Le Maguer, Naomi Harte, Simon King
Main reference Sébastien Le Maguer, Simon King, Naomi Harte: "The limits of the Mean Opinion Score for speech synthesis evaluation", Comput. Speech Lang., Vol. 84, p. 101577, 2024.
URL https://doi.org/10.1016/J.CSL.2023.101577

Speech synthesis has reached unprecedented quality, but its evaluation relies on methodologies developed more than two decades ago. Among these protocols, the Mean Opinion Score (MOS) test remains the overwhelming used standard. While not without controversy, and as contemporary synthesis systems produce speech remarkably close to human speech, it is now vital to determine how reliable this score is.

In this talk, I will present a series of four experiments to question MOS. With these experiments we address the following questions: How stable is the MOS of a system across time? How do the scores of lower quality systems influence the MOS of higher quality systems? How does the introduction of modern technologies influence the scores of past systems? How does the MOS of modern technologies evolve in isolation?

The outcome of our experiments confirms the relative nature of MOS. It also suggest that we may have reached the end of a cul-de-sac with current evaluation methodologies and that we are in critical need to develop more suited protocols.

4 Working groups

4.1 Day 1, session 1a. The people – collected personal statements

Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), Elisabeth André (Universität Augsburg, DE), Gérard Bailly (University Grenoble Alpes, FR), Erica Cooper (NICT – Kyoto, JP), Benjamin Cowan (University College – Dublin, IE), Naomi Harte (Trinity College Dublin, IE), Simon King (University of Edinburgh, GB), Esther Klabbers (Beaverton, US), Sébastien Le Maguer (University of Helsinki, FI), Zofia Malisz (KTH Royal Institute of Technology – Stockholm, SE), Bernd Möbius (Universität des Saarlandes, DE), Sebastian Möller (TU Berlin, DE & DFKI Berlin, DE), Roger K. Moore (University of Sheffield, GB), Ayushi Pandey (Trinity College Dublin, IE), Olivier Perrotin (University Grenoble Alpes, FR), Fritz Michael Seebauer (Universität Bielefeld, DE), Sofia Strömbergsson (Karolinska Institute – Stockholm, SE), Christina Tånnander (Swedish Agency for Accessible Media – Malmö, SE), David R. Traum (USC – Playa Vista, US), Petra Wagner (Universität Bielefeld, DE), Junichi Yamagishi (National Institute of Informatics – Tokyo, JP), and Yusuke Yasuda (Nagoya University, JP)

Since this seminar was designed to gather and reconcile as much as possible of the collective experiences in speech and speech synthesis evaluation, with few presentations and much discussion and collaborative work, we include all personal statements in this report, as they give a picture of the diversity of the issues at hand.

4.2 Day 2, session 1. Use case: Speech variation and training

Naomi Harte (Trinity College Dublin, IE), Fritz Michael Seebauer (Universität Bielefeld, DE), and Sofia Strömbergsson (Karolinska Institute – Stockholm, SE)

A shared feature of the scenarios here can be summarized as the following challenge:

What is the difference/distance between a given exemplar (what a speaker/TTS system is currently able to produce), and a currently un-realized target exemplar (what a speaker/TTS system is supposed to achieve).

To complicate things, there are a **multitude of potential target-appropriate exemplars**; in other words, there is a range of variation that is acceptable among exemplars meeting the requirements of "having achieved the target".

Evaluation of speech comes into play in two distinct but closely related ways here. We want to use TTS to generate an appropriate set of exemplars for the given user and scenario. This requires that in the absence of the ideal reference, we can assess whether a generated exemplar meets the required standard. This could be how e.g. certain vowels should be realised in the low-resource language, or how the voice of the child with a specific speech order should progress. The second way evaluation occurs is when, given the exemplars,

we then compare those to what is currently produced by the speaker/system. Within an acceptable level of variation, we need to judge how well the speaker/TTS is approaching production levels in the exemplars.

Thus a shared feature of the scenarios is that the **evaluation requires some kind of expert knowledge**; for TTS in a low-resource language, the expert knowledge required is being a native (native-like?) speaker of the target language. The evaluation of whether a speaker with a speech disorder or a speaker learning an L2 has reached (or is approaching) the target, requires expert listeners like an SLP or a language teacher, respectively. To achieve automatic evaluation therefore requires that we somehow capture what the expert would look out for, and embed that into the evaluation protocol. We believe this part is a major challenge to get right, and also essential.

Presumably, **the evaluation criteria will rely on expert-relevant parameters** referring to phonetics, phonology, rhythm, intonation, etc.

The **evaluation task for the expert is highly context-dependent** – the expert evaluator might, for example, need to know the age of the speaker (and might have a more lenient criteria for what is a target-appropriate exemplar for a younger speaker/child voice than for an older speaker). This is needed because the target production of the use cases depends on the speakers current level of capability which needs to be assessed. Thus we have a situation of a **moving target, as our required output or expectation of what can be produced is changing with time**.

4.3 Day 2, session 1. Use case: Speech-to-speech

Simon King (University of Edinburgh, GB), Sébastien Le Maguer (University of Helsinki, FI), Sebastian Möller (TU Berlin, DE & DFKI Berlin, DE), and Junichi Yamagishi (National Institute of Informatics – Tokyo, JP)

In addition to the production of speech using a different input material (e.g., text), a various set of applications requires to produce a target speech signal given a source speech signal. Among these applications, we considered three use cases: video dubbing, voice privacy, audio-only translation and simultaneous interpretation. Video dubbing consist of introducing a new audio track to a video. The most well known example is the adaptation of a movie to a new language. At the opposite of the audio-only translation, which consists of transforming a speech signal from one language to another language (e.g., podcast, radio program), this also implies to ensure a consistency between the video and the new audio track. Closely related is the simultaneous interpretation which consists of creating a speech signal in real time. A concrete example of application is when large summits are hold in multiple languages such as the UN or the EU assemblies. Finally, voice privacy differs from the previous use-cases. It requires the exact content of the message carried by the source speech signal to be transferred to the target signal. This should be done by ensuring that the speaker identity of the source signal is stripped from the target signal. All these use cases, except voice privacy, do not require speech technology and can be achieved using human speakers (e.g., voice actors for dubbing).

Each of these use cases has its own challenges. For example, it is imperative that part of the speaker identity is preserved for video dubbing, while for voiced privacy this identity should be stripped. Similarly, in the context of an audio only translation the speech-to-speech conversion can be achieved offline, while in the context of simultaneous interpretation the latency of the system will have an inpact on the comprehensibility of the interpretation. Even if we consider a system to be developped only for one use case, a "one fit all" evaluation protocol remains out of reach. For example, when producing a speech signal for dubbing, not only the speech signal has to be intelligibile but the synchronicity between the audio and the video should be preserved.

In order to design an adequate evaluation protocol, it will be important to determine which criteria have to be evaluated for each use case. During the working group, we identified six main criteria: the target speaker characteristics, the content preservation, the timing sensitivity, the expressivity of the speech, the real-time factor, and the impact of the non-speech acoustic information from the source signal. For each combination criterion/use case, different factors needs to be ensured. Table 1 summarizes the factors that we determined during the working group. This summary is non-exhaustive and should be seen as a starting point to determine what are the main points of evaluation.

There is a set of use cases where the source input is speech and the target output is speech. Source and target might be in the same or different languages. Some attributes, such as speaker identity, might be retained from source to target, or intentionally distinct. We considered dubbing videos, voice privacy, and audio-only translation. All of these can be achieved using natural speech, without technology. Applying speech technologies such as TTS or voice conversion was of course initially motivated by reducing cost. But technology has other advantages, including a lower barrier to entry (e.g., by reducing the skill level needed to produce a dub), increased speed/throughput of production, and a potentially wider portfolio of voices.

Table 1 Summary of factors to be evaluating when conducting an evaluation campaign dedicated to speech-to-speech use cases. All cells in bold correspond to critical point to be evaluated. If these are not assessed, the evaluation will be considered invalid.

CriterionUse Case	**Dubbing**	**Voice privacy**	**Translation**	**Interpretation**
Target speaker characteristics	Preserve Speaker ID (or) Fit to actor (or) **Consistency**	**Speaker ID masked**	Chosen by content producer	
Content	**Semantic fit** Artistic fit	Emotion recognition ER	**Semantic fit**	**Semantic fit**
Timing	**No overlap** Sync to video, lip sync			Ideal timing
Expressivity	Analogue to the source	Preserve within language **Mask speaker expressivity**	Analogue to the source	Neutrality
Non-Speech audio	Keeping other information Possibly diarization performance Noise robustness	Noise robustness	noise robustness	noise robustness
Real-Time Factor	offline	offline or online	offline	**online**

4.4 Day 2, session 1. Use case: Lengthy materials read aloud

Esther Klabbers (Beaverton, US), Erica Cooper (NICT – Kyoto, JP), Christina Tånnander (Swedish Agency for Accessible Media – Malmö, SE), and Yusuke Yasuda (Nagoya University, JP)

In Merriam-Webster, long-form is defined as "notably long in form in comparison to what is common or typical for works or content of a particular category". Since most TTS-generated outputs consists of single utterances or short sentences, "long-form" in a TTS context would include anything longer than a couple of sentences, for example a paragraph, news article or an entire book. Consequently, speech being evaluated need to have some degree of external validity to reflect a real-world reading/listening situation. Our discussions of what aspects of the speech to evaluate included the following topics:

Consistency and variability: The read speech should maintain consistency in audio quality, speaking rate and speech styles, as well as in the pronunciations of for example proper names and terms throughout the text. At the same time, the speech needs to be prosodic variation is crucial to ensure the speech remains comprehensibility and avoids causing listening fatigue.

Text contact: Whether delivered by a human narrator or TTS, the reader should exhibit what we here refer to as text contact: a sense of genuinely understanding the content being read.

Read dialogues: In fictional works, dialogue is a common feauture. It is essential to appropriately signal transitions between narrative text and dialogue, as well as between different characters within the dialogue.

4.5 Day 2, session 2. Methods

Sébastien Le Maguer (University of Helsinki, FI), Elisabeth André (Universität Augsburg, DE), Gérard Bailly (University Grenoble Alpes, FR), Erica Cooper (NICT – Kyoto, JP), Benjamin Cowan (University College – Dublin, IE), Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), Naomi Harte (Trinity College Dublin, IE), Simon King (University of Edinburgh, GB), Esther Klabbers (Beaverton, US), Zofia Malisz (KTH Royal Institute of Technology – Stockholm, SE), Bernd Möbius (Universität des Saarlandes, DE), Sebastian Möller (TU Berlin, DE & DFKI Berlin, DE), Roger K. Moore (University of Sheffield, GB), Ayushi Pandey (Trinity College Dublin, IE), Olivier Perrotin (University Grenoble Alpes, FR), Fritz Michael Seebauer (Universität Bielefeld, DE), Sofia Strömbergsson (Karolinska Institute – Stockholm, SE), Christina Tånnander (Swedish Agency for Accessible Media – Malmö, SE), David R. Traum (USC – Playa Vista, US), Petra Wagner (Universität Bielefeld, DE), Junichi Yamagishi (National Institute of Informatics – Tokyo, JP), and Yusuke Yasuda (Nagoya University, JP)

4.5.1 Best Practices

This group was interested in coming up with meta-guidelines to help researchers choose and develop evaluation methods for speech synthesis applications. One immediate challenge, lies in the fact that evaluation needs often, if not always, be tailored towards the use case it is supposed to measure. Following this, the guidelines outlined here need to be rather general, to still be applicable for all possible use case evaluations. We structure this abstracts in two different parts: First there is a non-exhaustive list of potential recommendations and second is a list of suggestions how to provide resources that aid researchers in implementing them. In listing the recommendations, we do not restrict ourselves by catering towards implementation problems of disseminating them in specific communities, but rather aim to provide a topline. The first guideline refers to the statistical validity of proposed methods. This mirrors the provided guidelines in adjacent fields like psychology. Examples for criteria to consider include adequate sample sizes for the number of investigated factors. The next point also contains scientific standards as a whole. For example the order randomisation of stimuli, the choice of adequate baseline references and control conditions, or the general fit of experiment design. We also consider that a plea towards open science would be sensible, suggesting the open sourcing of evaluation material and analysis code. In a less general recommendations we conclude that if possible, the use case of a given system should be considered when choosing or designing an evaluation protocol. On that same note, language considerations play into adapting already existing scales. These are findings borrowed from the field of psychometric research, that concepts might not be translateable between cultures and languages, and thusly scales would have to be re-developed or adapted if used in other contexts. This extends to older scales, that were designed to be valid for specific distributions of stimuli and might either be adapted for newer systems, or statistically corrected in the analysis phase. Most of these recommendations are geared towards individuals or institutions that have the time and resources to concern themselves with the pitfalls of designing proper evaluation protocols. Admitting that this is not always the reality of scientific research our last suggestion would be to consider collaboration or outsourcing the evaluation to ensure a certain quality standard is being kept. Moving into the topic of how to provide resources for

aspiring TTS researchers to use, when deciding to evaluate their systems. We consider using a linear, or branched, web-form template which instructs the user with guided questions akin to a decision tree. A similar resource could be a database which contains examples of different evaluation kinds and reference papers of confirmed quality, that could be queried by researchers looking for guidance. In terms of community work, we propose to somehow identify and showcase particularly well executed examples of evaluation, for example in form of an award. A more prescriptive approach would see the requirement of a standard in an evaluation section that should fulfill specific criteria, analogous to the well established standards of having a background in an introduction. Finally we consider what the specific meta-criteria might be to determine these "good" examples. It might be the thoroughness of the reporting done on all aspects of the decision making process, or the inventiveness with regards to already established methods.

4.5.2 Meta-structure

Starting from the use cases and the evaluation proposals. How can these be characterised, categorised?

Can the proposed evaluation be used for other use cases with no or little changes? Which ones? What do they have in common?

How do other evaluations fall within the same system?

Since I'm the only person turning up for this Group, I'm taking the liberty of promoting the evaluation "meta" taxonomy I've used in the past and which underpinned the EAGLES standards and resources activities coupled with more recent characterisation of the speech technology field ...

In particular, I suggest that is useful to partition use-cases into four broad categories according to whether a given scenario involves interactivity and/or requires real-time processing. For example, in a two-dimensional plot, voice-enabled artefacts would fall in the interactive/real-time quadrant, automated announcements would fall in the non-interactive/real-time quadrant, film dubbing would fall in the non-interactive/non-real-time quadrant, and long-form reading would fall in the interactive/non-real-time quadrant. (Note: I have a LaTeX template suitable for this diagram – slide 4 in my presentation.)

For real-time/interactive use cases, I also recommend distinguishing between three behavioural domains:

- the physical domain of objects and actions (for mechanical support),
- the information domain of knowledge and data (for cognitive support),
- the social domain of agents and relations (for emotional support).

Interaction in the physical and information domains typically involves formulaic speech acts – "command-and-control" or "question-and-answer" respectively – which usually conform to a strict "turn-taking" protocol for dialogue. Interaction in the social domain involves more fluid conversational behaviour with considerable overlap between interlocutors. These domains are not mutually-exclusive, hence any particular use-case will have a balance of requirements across all three domains. (Note: I have a diagram that captures this – slide 5 in my presentation.)

Of course, there are many factors which influence the ultimate performance of spoken language systems. This means that, not only is it is necessary to distinguish between "capabilities" and "requirements" of the components technologies (such as speech synthesis), but it also important to emphasise that the purpose of introducing spoken language technology into an application is to achieve the appropriate operational benefits. However, successful

implementation of spoken language systems depends only indirectly on the technical features of the system components. The anticipated application benefits need to be expressed in terms of application requirements which in turn need to be expressed as technical requirements. On the other side of the coin, the features of the technology need to be expressed in terms of technical capabilities which in turn need to be expressed as application capabilities. Both the technical and application capabilities/requirements are multi-dimensional in nature and thus require assessment of "goodness" across a set of relevant application and technical factors. (Note: I have a diagram that captures this – slide 6 in my presentation.)

Finally, it is important to acknowledge that a key "goodness" parameter (alongside obvious dimensions such as intelligibility) is the "appropriateness" of a particular synthetic voice to a given communicative context. In this regard it is useful to distinguish the (dynamic) situational context from the (static) embodied context. The latter would likely be motivated by suitable design principles/priors, and thus evaluation (and therefore, optimisation) could be performed off-line. The former implies synthesis that is reactive to changing conditions, and thus evaluation (and therefore, optimisation) would require on-line evaluation.

4.5.3 Users and user expectations

An important consideration for the evaluation of speech synthesis systems is the selection of a set of human raters that is representative for the population of end users of the system. Options range from drawing random samples from the general population if the target application of the system is a general-purpose speech synthesis system, to a system that serves the need of users with specific properties or needs. Examples for the latter are people with physiological (e.g., vision, hearing, reading) or neurological impairments (e.g., aphasia, personality disorders, people who would benefit from Easy Language). For these more specific user groups, coverage by a randomly drawn set of human evaluators does not appear to be feasible. A perfect match between evaluators and target users is required for end users with cognitive or neurological impairments. Furthermore, developers of synthesis systems are interested in diagnostic information about deficiencies in the performance of their systems. While the developers themselves should be excluded from independent evaluation efforts involving their own systems, some degree of expertise in speech science and technology is required to identify errors in the synthesis output. That said, the paradigm known as "yuck detection" has recently been applied with some success: evaluators hit a button whenever they perceive a flaw in the synthesized signal, which allows system developers to inspect locations with accumulated indications of flaws. End users may also have different expectations of the synthesis quality and the system's capabilities, which may introduce judgment biases. A synthesis system that achieves near-human-likeness in overall quality and intelligibility may be expected to be also capable of maintaining a dialog, manage turn-taking, and perhaps even be omniscient, even though such capabilities are beyond the scope of the speech synthesizer itself. Moreover, it is presently unclear how users cope with inconsistent, apparently stochastic synthesis output resulting from the statistical nature of state-of-the-art speech generation methods. In summary, depending on the aim of the evaluation and the application domain of speech synthesis, designing subjective evaluation setups needs to make judicious decisions about the composition of the pool of human synthesis evaluators.

4.5.4 Pitfalls and problems

There are some common problems and pitfalls that apply to all evaluation protocols and for all use cases. These can be in the design of the protocol, during its implementation and execution, or when interpreting the results.

A typical design pitfall is failing to adequately simulate the intended use case and thus create an ecologically-invalid design. The two principal errors made in implementation are to use poorly-chosen materials or unsuitable participants (or objective measures).

Examples of poor design: evaluating speech as audio-only when the use case is a talking robot; ...

Examples of poorly-chosen materials: including incomplete or ungrammatical sentences in the test set; using sentences from out-of-copyright novels when the intended use case is spoken dialogue.

Examples of unsuitable participants: listeners who are unfamiliar with the accent of the synthetic speech and so are unable to accurately judge speaker similarity; ...

Examples of incorrect results interpretation: ...

Design

- Use Case 1:
  - side effects also need to be assessed
- Use Case 2:
  - how to determine the right evaluator
  - is the listener qualified to evaluate the sample
  - resource issue
- Use Case 3:
  - over sensitivity of audio only evaluation in case of multimodal (some artefacts may not matter)
- Use Case 4:
  - user expectation/familiarity needs to be taken into account
  - difference of meaning/implications between same terminology (e.g., what is "fair"?)
- Use Case 5:
  - assuming traditional way of doing the evaluation is the gold standard ⇒ metrics should not become the target
  - over sensitivity of audio only evaluation in case of multimodal (some artefacts may not matter)

Execution

- Use Case 1:
  -
- Use Case 2:
  - different experience between observer and user (⇒ overthinking from the observer perspective, influence of the recovery)
  - user vs participant
- Use Case 3:
  - what if the system if not fast enough yet ⇒ offline, how to deal with fine grain
- Use Case 4:
  - input processing
  - no access to the final use ⇒ defining the delta
  - "inappropriate" material deployed during the test
  - proficiency of the participant in the language
  - cognitive load (dual language – evaluation/evaluated language)

- Use Case 6:
  - what about participants who don't usually listen to audiobooks?
  - expectations from participant how have already read the book?
  - duration tolerance (5min is ok but 2h) + long term consistency

Interpretation

- Use Case 1:
  - what if basic test works only for formant synthesis?
- Use Case 3:
  - over interpretation of the results

4.6 Notes on proposed methods and method requirements

4.6.1 Suggested evaluation methods for some use cases 1

Use case	Needs of end users	Metric
Designing for an artefact	aligned multimodal interaction affordances	positioning of persona (within space of possible voices)
Targeted manipulation in a high-quality voice	disentanglement of effects within an experimental setting	exact reproduction of intended contrast
	preservation of original "voice"	objective or subjective analysis of signal degradation and/or speaker ID

4.6.2 Suggested evaluation methods for some use cases 2

1. Virtual human coach- visual and tactile; social context- weird aspects of imitation; derogatory aspects of voice mimicry?
2. Augmented Accessibility- Identity uniqueness; creating voice that satisfies constraints; unique characteristics of the conditions; age appropriateness; speech loss; representation and social relationship
3. Agent/Robot mediator

Metrics/Observations

1. Appropriateness; social acceptability; trust; representation matching
2. Appropriateness; social acceptability; social and dialogue cues; temporal aspects of interaction; qualitative identification of dimensions of voice key to the user (interviews); communication behaviours (turns etc); identity matching/augmentation; consistency; configurability, satisfaction with result

Experiment Design

1. Third party vs user/creator evaluation; Explicit- asking them; Implicit – Engaging and interacting Comparators? control conditions?

4.6.3 Suggested evaluation methods for some use cases 3

Criterion/Use case	**Dubbing (Audio/Visual)**	**Voice privacy/Anonymization**	**Translation (audio only, e.g podcast)**	**Simultaneous interpretation (e.g., EU parliament, meeting)**
Target speaker characteristics	Speaker ID Speaker embedding Fit to actor Consistency	Speaker ID Speaker embedding	Chosen by content producer	Don't care
Content	AVSR/ASR WER Artistic and semantic fit	ASR WER Emotion recognition ER	BERT score Semantic fit (e.g., journalistic fit)	BERT score Semantic fit (e.g., journalistic fit)
Timing	Must have :: No overlap Nice to have :: Sync to video, lip sync			When to speak? (turn taking, cognitive load, latency)
Prosody	Expressivity target being the "same" as the source	Expressivity same (within language, without speaker)	Expressivity target being "same" as the source	Neutrality
Non-speaker related information in the source (channel, music ...)	Keeping other information Possibly diarization performance Noise robustness	Noise robustness		
Offline / Online	offline (real time factor)	offline / online	offline	oneline

Evaluation methods. Long-form speech should primarily be evaluated by humans, although automatic methods can be used in some cases.

We identified two distinct evaluation scenarios: (1) general evaluation of a TTS voice developed for long-form content; and (2) evaluation of the final product, such as an audiobook.

In the second scenario, automatic methods such as **ASR** (Automatic Speech Recognition) can compare its output with the input to the TTS; **LLMs** (Large Language Models) can analyse linguistic complexity to identify anomalies; and a silence detector can identify pauses that are too long, which might occur when a human narrator loses focus.

In our selected use case, the shifts between narrative text and dialogue and between speakers, we propose using **ARS** (Audience Response System), where the respondents push a button whenever they perceive a shift. This procedure is followed up with questions about perceived difficulty. Similar methods could evaluate whether the listeners can identify the speaker in multi-party dialogues, for instance by clicking a picture of the speaker.

Finally, we agreed that the test data selection is crucial and must involve challenging text passages to thoroughly assess the system's capabilities.

4.7 Day 2, session 1. Use case: Simulation and stimuli generation (for speech science)

Zofia Malisz (KTH Royal Institute of Technology – Stockholm, SE), Roger K. Moore (University of Sheffield, GB), and Petra Wagner (Universität Bielefeld, DE)

This group discussed quality dimensions and evaluation metrics in which synthesis (or the analysis of synthetic speech) is used with the expressed reason to conduct scientific research. Perhaps the most typical case for this will be the application of synthesis within the area of speech science, phonetics or related fields, but it may also include other disciplines in which the vocalizations and their quality are concerned. Within the group, two concrete use cases were discussed and further operationalized:

Use Case 1: Manipulation of voice alongside a certain dimension of scientific interest (f0, prominence, age, manner or place of articulation etc.)

Use Case 2: Designing a voice for an artefact (that is not normally considered to have a voice)

For Use Case 1, we identified two core dimensions of quality, which are known challenges for state-of-the-art synthesis systems:

(1) the successful disentanglement of effects within a speech science experimental setting. This could be operationalized by measuring the exact reproduction of the intended contrast encoded by this dimension (e.g., pitch modification). However, a truly successful manipulation would necessarily also check that the manipulation does not result in an unintended modification of those features that characterize the original "voice", or speaker identity, e.g., by not only changing pitch, but also the perceived gender.

(2) Thus, the "preservation of original voice" is our second quality dimension. This could be operationalized in several ways, but example metrics would be an objective analysis of signal degradation and the preservation of the speaker ID. Another metric would be subjective listening tests to perform these analyses.

For Use Case 2, we identified two core dimensions of quality:

(1) aligned multimodal interaction affordances, which could be evaluated with evaluation methods that have been established in HCI, including the alignment with the user, ventriloquist effect, or general usability of the artefact. A second dimension of quality would be (2) the adequate "positioning of the persona" established by an artefact's voice. This could be measured by perceived appropriateness or the artefact's voice, or rather, the voice's congruency with its perceived embodiment. An example for this would be that a "speaking briefcase" would probably be expected to have a voice that is somewhat muffled (due to the fabric it consists of), and has a voice that is aligned with its size.

During our discussions it emerged that our two use cases share one quality dimension, which is the "appropriateness, or congruency" of a generated voice with the persona it belongs to. If extended, this quality dimension could also define the boundary between human-like and "super-humanlike" voices, or between different speaker groups or even species.

4.8 Day 2, session 1. Use case: Incremental TTS; incremental speech (i.e. live speech production)

Bernd Möbius (Universität des Saarlandes, DE), Gérard Bailly (University Grenoble Alpes, FR), Jens Edlund (KTH Royal Institute of Technology – Stockholm, SE), and Ayushi Pandey (Trinity College Dublin, IE)

We're defining these systems loosely as signal generation systems (i.e. a TTS engines) meeting the following criteria:

Condition the signal on a stream of tokens that is smaller than what is common.

If "sentence" or "utterance" is the typical unit size for TTS; we may use lemma/word, grapheme, or phoneme.

1. Leave open a continuous input channel representing external phenomena (e.g. surrounding noise)
2. Leave open a continuous input channel representing phenomena that are conditioned or created by self (e.g. own audio, reactions of others).
3. Be able to take decisions, midstream, based on these extra input channels. It may pay off to view the standard conditioning stream as intention, and the other streams as controls.

Since this starts with a capability description, in quite technical terms, it's perfectly feasible to have a feature assessment – checkboxes pretty much – that must be passed in order for a system to count. That narrows down our concept space neatly.

Also, this can be seen as a component. In which case API descriptions can be used as capability descriptors, and a clear eval step is to see if the effect of using the API is as follows. So claimed capabilities and capabilities delivered.

As with embodied/virtual agents, more of a technology than a use case (and there are of course relevant pure technology evaluations). Use cases include anything interactional that isn't heavily turn based, as well as fast and responsive transmission. Incremental rendering is also at least in principle less resource (memory) consuming, And a prerequisite for any kind of situational real-time adaptation of the speech, such as the Lombard effect.

- Interruptible systems
- Realtime adaptation (e.g. Lombard)
- Responsive speech
- Listening speakers
- Feedback-sensitive speech
- Self-correcting speech

Good use cases may include: listening speakers (interruptible, feedback responsive), self-corrective speech, Lombard speech. And possibly streaming speech.

4.9 Day 2, session 1. Use case: (Human-like) embodied spoken dialogue

Petra Wagner (Universität Bielefeld, DE), Elisabeth André (Universität Augsburg, DE), Benjamin Cowan (University College – Dublin, IE), and Olivier Perrotin (University Grenoble Alpes, FR)

Our group conducted a deep discussion on the use and evaluation of synthesis when applied embodied speech systems, i.e., where a speech synthesiser is embodied either within an external agent (virtual or robot) or a human user (in the case of Assistive Augmented Communication (AAC) interactions). Overall, we could not identify gold standards for the evaluation of such systems, which go largely beyond clarity and intelligibility. Evaluation is extremely task-specific and so are the associated metrics. Instead, we encourage the definition of gold standard process for selecting the evaluation practices and associated metrics rather than to define them. We identified two main goals of such embodied system:

1. establish an identity;
2. being a dialogue partner in a social environment.

As for identity, human likeness which is highly prized in most TTS applications is not necessarily a requirement in this use case, for both agent and patient voice embodiment. By contrast, consistency being voices and their body is crucial (as opposed to have several agents with similar voices or vice-versa) to allow distinctiveness and recognition of agents/patients as individual entities. Also, identification of voice might be part of a brand identity. Plausibility is also a criteria, i.e. the voice matches the embodiment, to favour expectation matching from interlocutor. Having identified those criteria, two complementary evaluation methods are relevant. On the one hand, the use of explicit detailed questionnaires addressed to both participants in interaction with the system under evaluation, and external participants watching the interaction. On the other hand, the getting of implicit global preference about interaction experiences with several voices. Finally, performing these evaluations longitudinally is a mean to integrate the familiarisation of the participants with the new voice to evaluate.

The second direction is the assessment of the interaction, mainly being whether the message is conveyed properly (either via linguistic and/or paralinguistic information), and whether the voice impact user language and dialogue. A variety of paradigms have been listed for evaluation, transversally of the task-specificity of interactions. First, as for identity assessment, the evaluation can be carried out either by extracting explicit metrics from the speech signals or questionnaires, or by inferring implicit metrics from the interlocutor reaction to the speech interaction. In the former case, borrowing metrics from voice coaching would be an interesting direction to investigate. In the latter case, the interlocutor has the role of feature extractor, which is highly dependant on his/her social background and environment. These approaches can be carried out offline (post experiment), or online. In that case, in a similar fashion that the yuck test, the laugh test naturally measures unexpected situations. The choice of evaluators is also of high importance, as evaluation from the participants involved in the interaction and the one of external viewers should be quite complementary. In the latter case, the question of realistic immersion in the interaction can have great impact on the assessment. Finally, the quantifying of user-agent interplay such as entrainment/adaptation to voice/linguistic markers alignment could be one aspect of assessing interaction, but is highly task- and environment-dependent to assume whether alignment is an expected or unexpected phenomena to happen in a statisfying interaction.

Participants

- Elisabeth André
Universität Augsburg, DE
- Gérard Bailly
University Grenoble Alpes, FR
- Erica Cooper
NICT – Kyoto, JP
- Benjamin Cowan
University College – Dublin, IE
- Jens Edlund
KTH Royal Institute of Technology – Stockholm, SE
- Naomi Harte
Trinity College Dublin, IE
- Simon King
University of Edinburgh, GB
- Esther Klabbers
Beaverton, US
- Sébastien Le Maguer
University of Helsinki, FI
- Zofia Malisz
KTH Royal Institute of Technology – Stockholm, SE
- Bernd Möbius
Universität des Saarlandes – Saarbrücken, DE
- Sebastian Möller
TU Berlin, DE & DFKI Berlin, DE
- Roger K. Moore
University of Sheffield, GB
- Ayushi Pandey
Trinity College Dublin, IE
- Olivier Perrotin
University Grenoble Alpes, FR
- Fritz Michael Seebauer
Universität Bielefeld, DE
- Sofia Strömbergsson
Karolinska Institute – Stockholm, SE
- Christina Tånnander
Swedish Agency for Accessible Media – Malmö, SE
- David R. Traum
USC – Playa Vista, US
- Petra Wagner
Universität Bielefeld, DE
- Junichi Yamagishi
National Institute of Informatics – Tokyo, JP
- Yusuke Yasuda
Nagoya University, JP

Report from Dagstuhl Seminar 25041

Solving Problems on Graphs: From Structure to Algorithms

Akanksha Agrawal[*1], Maria Chudnovsky[*2], Daniël Paulusma[*3], Oliver Schaudt[*4], and Julien Codsi[†5]

1 **Indian Institute of Techology Madras, IN.** akanksha.agrawal.2029@gmail.com
2 **Princeton University, US.** mchudnov@math.princeton.edu
3 **Durham University, GB.** daniel.paulusma@durham.ac.uk
4 **Bayer AG – Leverkusen, DE.** oliver.schaudt@bayer.com
5 **Princeton University, US.** jc3530@princeton.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25041 "Solving Problems on Graphs: From Structure to Algorithms", which was held from 19 January to 24 January 2025. The report contains abstracts for presentations about recent structural and algorithmic developments for a variety of graph problems. It also contains a collection of open problems which were posed during the seminar.

Seminar January 19–24, 2025 – https://www.dagstuhl.de/25041
2012 ACM Subject Classification Theory of computation → Graph algorithms analysis; Theory of computation → Fixed parameter tractability; Mathematics of computing → Graph algorithms; Theory of computation → Problems, reductions and completeness
Keywords and phrases computational complexity, graph algorithms, graph classes, graph containment relations, graph width parameters
Digital Object Identifier 10.4230/DagRep.15.1.105

1 Executive Summary

Akanksha Agrawal (Indian Institute of Techology Madras, IN)
Maria Chudnovsky (Princeton University, US)
Daniël Paulusma (Durham University, GB)
Oliver Schaudt (Bayer AG – Leverkusen, DE)

Many discrete optimization problems can be modeled as graph problems, leading to a long list of well-studied problems, which include graph partitioning, covering and packing problems, network design problems, width parameter problems, and so on. Most of these graph problems are computationally hard. However, this situation may change if we require the input to belong to some special graph class. This leads to two fundamental questions, which formed the focus of our Dagstuhl Seminar: for which classes of graphs can a computationally hard graph problem be solved in polynomial time, and for which classes of graphs does the problem remain hard?

One of the main research aims of our seminar was to discover new insights that lead to results for a whole range of problems rather than just for a single problem alone. That is, our goal was to determine general properties, such that large classes of graph problems sharing certain common features can be solved efficiently on a graph class if and only if the graph

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

class has these properties. For this purpose, our seminar brought together researchers from Discrete Mathematics, working in structural graph theory, and researchers from Theoretical Computer Science, working in algorithmic graph theory. In total, 41 participants from 13 different countries attended the seminar.

The scientific program of the seminar consisted of 18 sessions: 5 survey talks of fifty minutes, 10 contributed talks of at most thirty minutes, and 5 open problem sessions. This left ample time for discussions and problem-solving. Participants presented the progress that was made during the seminar during several "progress report" sessions.

Each of the five survey talks covered a particular structural or algorithmic key aspect of the seminar in order to enable collaborations of researchers with different backgrounds. On Monday, Tuukka Korhonen presented a survey on new results on induced minors of graphs. This is a well-known graph containment relation, but recently developed techniques led to significant progress and new open problems. On the same day, Édouard Bonnet discussed a number of new results and open problems on twin-width, a relatively new graph width parameter that is being studied intensively. The final survey talk on Monday was given by Stefan Kratsch, who discussed a number of open problems around modular-based graph parameters and graph width parameters from the perspective of parameterized complexity.

On Tuesday, Maria Chudnovsky presented an overview of the induced subgraphs and tree decompositions project that currently comprises a series of 18 papers. Afterward, Nicolas Trotignon explained the importance of layered wheels. These are graphs of arbitrarily large treewidth that play a significant role in the induced subgraphs and tree decompositions project. In general, a layered wheel consists of many paths, all pairwise with edges between them (thus guaranteeing large treewidth), where many additional desired properties may be forced. Thus layered wheels provide a useful source of boundary examples of families of graphs with large treewidth. A few weeks after the completion of our seminar, three of the participants, Bogdan Alecu, Édouard Bonnet, and Nicolas Trotignon, found, together with Pedro Bureo Villafana, a large new family of layered wheels providing counterexamples to several known conjectures in the field, some of which were discussed at the seminar.

The five general open problem sessions took place on Monday, Tuesday, Wednesday, and Thursday. Details of the presented problems can be found in the report, together with abstracts of all the talks.

We thank Julien Codsi for his help with the Dagstuhl Report of our seminar.

2 Table of Contents

3 Overview of Talks

3.1 Twin-Width: Algorithmic Applications and Open Questions

Édouard Bonnet (ENS – Lyon, FR)

We review some algorithmic applications of the relatively new twin-width parameter, and propose several open problems.

3.2 Forbidding induced subgraphs: structure and algorithms

Maria Chudnovsky (Princeton University, US)

Tree decompositions are a powerful tool in both structural graph theory and graph algorithms. Many hard problems become tractable if the input graph is known to have a tree decomposition of bounded "width". Exhibiting a particular kind of a tree decomposition is also a useful way to describe the structure of a graph.

Tree decompositions have traditionally been used in the context of forbidden graph minors; studying them in connection with graph containment relations of more local flavor (such as induced subgraph or induced minors) is a relatively new research direction. In this talk we will discuss recent progress in this area, touching on both the classical notion of bounded tree-width, and concepts of more structural flavor.

In particular we will describe a recent result providing a complete list of induced subgraph obstructions to bounded pathwidth.

3.3 Tree independence number of graphs with no induced $S_{t,t,t}$, $K_{t,t}$ and line graph of subdivided walls

Julien Codsi (Princeton University, US), Maria Chudnovsky (Princeton University, US), Daniel Lokshtanov (University of California – Santa Barbara, US), Martin Milanič (University of Primorska, SI)

Joint work of Julien Codsi, Maria Chudnovsky, Daniel Lokshtanov, Martin Milanič, Varun Sivashankar
Main reference Maria Chudnovsky, Julien Codsi, Daniel Lokshtanov, Martin Milanic, Varun Sivashankar: "Tree independence number V. Walls and claws", CoRR, Vol. abs/2501.14658, 2025.
URL https://doi.org/10.48550/ARXIV.2501.14658

The tree independence number is a natural generalization of treewidth, offering versatility for a wide range of algorithmic applications. In this talk, we explore the reasoning behind a recent result that establishes a polylogarithmic bound on the tree independence number for graphs excluding certain structures: induced $K_{t,t}, S_{t,t,t}$ ($K_{1,3}$ where each edge is subdivided $t-1$ times), and the line graph of a subdivided large wall. Consequently, this graph class admits quasipolynomial-time algorithms for numerous problems that are typically NP-Hard. Beyond its algorithmic implications, our method provides valuable structural insights, particularly into the identification of desirable balanced separators within this graph class.

3.4 Bandwidth is FPT-Approximable

Daniel Lokshtanov (University of California – Santa Barbara, US), Maria Chudnovsky (Princeton University, US)

Joint work of Daniel Lokshtanov, Eran Nevo, Maria Chudnovsky

A linear layout of a graph G is a bijection $f : V(G) \to 1...n$, and the bandwidth of a linear layout f is the maximum over all edges uv of G of $|f(u) - f(v)|$. The bandwidth $bw(G)$ of a graph G is the minimum bandwidth of a linear layout of G.

Computing the bandwidth of an input graph G is a notoriously hard problem: it is NP-hard to approximate within any constant factor [1] and W[t] hard for every t [2] , and these hardness results carry over even to very restricted sub-classes of trees.

We prove that bandwidth is FPT-approximable: there exists an algorithm that takes as input a graph G and integer k, runs in time $f(k)n^{O(1)}$ and either concludes that the bandwidth of G is more than k, or produces a layout with bandwidth at most $g(k)$.

On the way we show the following structural result: there exists a function h, such that for every graph G and integer k, either G contains a sub-tree of bandwidth at least k, or the bandwidth of G is at most $h(k)$.

References

1 Chandan Dubey, Uriel Feige and Walter Unger. *Hardness results for approximating the bandwidth*. Journal of Computer and System Sciences, 2011, Volume 77, p.62-90.
2 Hans L. Bodlaender and Michael R. Fellows. *W[2]-hardness of precedence constrained K-processor scheduling*. Operations Research Letters, Volume 18, Issue 2, 1995

3.5 Graphs of bounded sphere dimension

Meike Hatzel (IBS – Daejeon, KR)

Joint work of Meike Hatzel, James Davies, Agelos Georgakopoulos, Rose McCarty

The *sphere dimension* of a graph G is the smallest integer $d \geq 2$ so that G is an intersection graph of metric spheres in $\mathbb{R}^d$. This talk considers the class $\mathcal{C}^d$ of graphs with sphere dimension d. We present the results that for each integer t, the class of all graphs in $\mathcal{C}^d$ that exclude $K_{t,t}$ as a subgraph has strongly sublinear separators and that $\mathcal{C}^d$ has asymptotic dimension at most $2d + 2$.

3.6 Minimal obstructions to C_5-coloring in hereditary graph classes

Jorik Jooken (KU Leuven, BE), Jan Goedgebeur (KU Leuven, BE), Paweł Rzążewski (Warsaw University of Technology, PL), and Oliver Schaudt (Bayer AG – Leverkusen, DE)

Joint work of Jorik Jooken, Jan Goedgebeur, Karolina Okrasa, Paweł Rzążewski, Oliver Schaudt
Main reference Jan Goedgebeur, Jorik Jooken, Karolina Okrasa, Paweł Rzążewski, Oliver Schaudt: "Minimal Obstructions to C_5-Coloring in Hereditary Graph Classes", in Proc. of the 49th International Symposium on Mathematical Foundations of Computer Science, MFCS 2024, August 26-30, 2024, Bratislava, Slovakia, LIPIcs, Vol. 306, pp. 55:1–55:15, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024.
URL https://doi.org/10.4230/LIPICS.MFCS.2024.55

For graphs G and H, an H-coloring of G is an edge-preserving mapping from $V(G)$ to $V(H)$. Note that if H is the triangle, then H-colorings are equivalent to 3-colorings. In this paper we are interested in the case that H is the five-vertex cycle C_5.

A minimal obstruction to C_5-coloring is a graph that does not have a C_5-coloring, but every proper induced subgraph thereof has a C_5-coloring. In this paper we are interested in minimal obstructions to C_5-coloring in F-free graphs, i.e., graphs that exclude some fixed graph F as an induced subgraph. Let P_t denote the path on t vertices, and let $S_{a,b,c}$ denote the graph obtained from paths $P_{a+1}, P_{b+1}, P_{c+1}$ by identifying one of their endvertices.

We show that there is only a finite number of minimal obstructions to C_5-coloring among F-free graphs, where $F \in \{P_8, S_{2,2,1}, S_{3,1,1}\}$ and explicitly determine all such obstructions. This extends the results of Kamiński and Pstrucha [1] who proved that there is only a finite number of P_7-free minimal obstructions to C_5-coloring, and of Dębski et al. [2] who showed that the triangle is the unique $S_{2,1,1}$-free minimal obstruction to C_5-coloring.

We complement our results with a construction of an infinite family of minimal obstructions to C_5-coloring, which are simultaneously P_{13}-free and $S_{2,2,2}$-free. We also discuss infinite families of F-free minimal obstructions to H-coloring for other graphs H.

References

1 Marcin Kamiński and Anna Pstrucha. *Certifying coloring algorithms for graphs without long induced paths*. Discrete Applied Mathematics, Volume 261, 2019, Pages 258-267
2 Michał Dębski, Zbigniew Lonc, Karolina Okrasa, Marta Piecyk, and Paweł Rzążewski. *Computing Homomorphisms in Hereditary Graph Classes: The Peculiar Case of the 5-Wheel and Graphs with No Long Claws*. In 33rd International Symposium on Algorithms and Computation (ISAAC 2022). Leibniz International Proceedings in Informatics (LIPIcs), Volume 248, pp. 14:1-14:16, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2022)

3.7 Induced minors of graphs

Tuukka Korhonen (University of Copenhagen, DK)

The theory of induced minors of graphs aims to generalize the theory of graph minors to dense graph classes. This dates back to the end of the 80s and the start of the 90s, when several negative results in this direction were proven. However, in the recent years there has been a resurgence of interest into induced minors, and some positive results and interesting conjectures have been discovered. In this talk I survey the theory of induced minors and the most interesting open problems in this area, focusing mostly on the algorithmic perspective.

3.8 Graph structure and parameterized algorithms

Stefan Kratsch (HU Berlin, DE)

Joint work of Narek Bojikian, Vera Chekan, Falko Hegerfeld, Stefan Kratsch, Pascal Kunz

The talk surveys some recent work on parameterized algorithms for graph problems relative to structural parameters: (1) Efficient parameterized algorithms that obtain in polynomial time a approximate solution with additive error depending on a chosen graph parameter. (2) Key ideas behind the recent tight bound for Steiner Tree parameterized by clique-width. (3) Tight bounds relative to multi-clique-width.

3.9 Restricted CSPs and H-free Algorithmics

Barnaby Martin (Durham University, GB)

Joint work of Barnaby Martin, Santiago Guzman-Pro

In recent years, much attention has be placed on the complexity of graph homomorphism problems when the input is restricted to P_k-(subgraph)-free graphs. We consider the directed version of this research line, by addressing the question: is it true that digraph homomorphism problems CSP(H) have a P versus NP-complete dichotomy when the input is restricted to DP_k-(subgraph)-free digraphs (where DP_k is the directed path on k vertices)? We build on the theory of constraint satisfaction problems to address this question.

Our first main results are a P versus NP-complete classification of CSPs when the input is restricted to $\mathcal{F}$-homomorphism-free digraphs, or restricted to CSP(H') for some finite digraph H'. We then use the established connection to constraint satisfaction theory to show our third main result (and partial answer to the question above): if CSP(H) is NP-complete, then there is a positive integer N such that CSP(H) remains NP-hard even for DP_N-(subgraph)-free digraphs. Moreover, CSP(H) remains NP-hard for DP_N-(subgraph)-free acyclic digraphs, and becomes polynomial-time solvable for DP_{N-1}-(subgraph)-free acyclic digraphs.

Another contribution of this work is verifying the question above for digraphs on three vertices and a family of smooth tournaments.

3.10 The Complexity of Diameter on H-free graphs

Jelle Oostveen (Utrecht University, NL)

Joint work of Jelle Oostveen, Daniël Paulusma, Erik Jan van Leeuwen
Main reference Jelle J. Oostveen, Daniël Paulusma, Erik Jan van Leeuwen: "The Complexity of Diameter on H-free Graphs", in Proc. of the Graph-Theoretic Concepts in Computer Science – 50th International Workshop, WG 2024, Gozd Martuljek, Slovenia, June 19-21, 2024, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 14760, pp. 444–459, Springer, 2024.
URL https://doi.org/10.1007/978-3-031-75409-8_31

The intensively studied DIAMETER problem is to find the diameter of a given connected graph. We investigate, for the first time in a structured manner, the complexity of DIAMETER for H-free graphs, that is, graphs that do not contain a fixed graph H as an induced subgraph. We first show that if H is not a linear forest with small components, then DIAMETER cannot be solved in subquadratic time for H-free graphs under SETH. For some small linear forests, we do show linear-time algorithms for solving DIAMETER. For other linear forests H, we make progress towards linear-time algorithms by considering specific diameter values. If H is a linear forest, the maximum value of the diameter of any graph in a connected H-free graph class is some constant $d_{\max}$ dependent only on H. We give linear-time algorithms for deciding if a connected H-free graph has diameter $d_{\max}$, for several linear forests H. In contrast, for one such linear forest H, DIAMETER cannot be solved in subquadratic time for H-free graphs under SETH. Moreover, we even show that, for several other linear forests H, one cannot decide in subquadratic time if a connected H-free graph has diameter $d_{\max}$ under SETH.

3.11 Partial Grundy Coloring is FPT

Fahad Panolan (University of Leeds, GB)

Joint work of Akanksha Agrawal, Daniel Lokshtanov, Fahad Panolan, Saket Saurabh, Shaily Verma
Main reference Akanksha Agrawal, Daniel Lokshtanov, Fahad Panolan, Saket Saurabh, Shaily Verma: "Parameterized Saga of First-Fit and Last-Fit Coloring", in Proc. of the 42nd International Symposium on Theoretical Aspects of Computer Science, STACS 2025, March 4-7, 2025, Jena, Germany, LIPIcs, Vol. 327, pp. 5:1–5:21, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025.
URL https://doi.org/10.4230/LIPICS.STACS.2025.5

The classic greedy coloring algorithm considers the vertices of an input graph G in a given order and assigns the first available color to each vertex v in G. In the Grundy Coloring problem, the task is to find an ordering of the vertices that will force the greedy algorithm to use as many colors as possible. In the Partial Grundy Coloring, the task is also to color the graph using as many colors as possible. This time, however, we may select both the ordering in which the vertices are considered and which color to assign the vertex. The only constraint is that the color assigned to a vertex v is a color previously used for another vertex if such a color is available. Aboulker et al. [1] proved that Grundy Coloring is W[1]-hard. In this talk, we prove that Partial Grundy Coloring is fixed-parameter tractable.

References

1 Pierre Aboulker, Édouard Bonnet, Eun Jung Kim, and Florian Sikora. *Grundy Coloring and Friends, Half-Graphs, Bicliques*. In 37th International Symposium on Theoretical Aspects of Computer Science (STACS 2020). Leibniz International Proceedings in Informatics (LIPIcs), Volume 154, pp. 58:1-58:18, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2020)

3.12 Erdős-Pósa property of tripods in directed graphs

Michal Pilipczuk (University of Warsaw, PL), Meike Hatzel (IBS – Daejeon, KR)

Joint work of Michal Pilipczuk, Meike Hatzel, Karolina Okrasa, Marcin Briański
Main reference Marcin Brianski, Meike Hatzel, Karolina Okrasa, Michal Pilipczuk: "Erdös-Pósa property of tripods in directed graphs", CoRR, Vol. abs/2408.16733, 2024.
URL https://doi.org/10.48550/ARXIV.2408.16733

In a directed graph D with sources S and sinks T, a tripod is the union of a pair of $S - T$ paths that have different sources but the same sink. We prove that tripods in directed graphs have the Erdős-Pósa property. The key element in the proof is the use of the Matroid Intersection Theorem. We will also mention open problems about possible generalizations and related algorithmic problems.

3.13 Maximum k-colorable Induced Subgraph (and beyond) in Polynomial Time

Roohani Sharma (MPI für Informatik – Saarbrücken, DE)

Joint work of Roohani Sharma, Akanksha Agrawal, Paloma T. Lima, Daniel Lokshtanov, Paweł Rzążewski, Saket Saurabh, Meirav Zehavi
Main reference Akanksha Agrawal, Paloma T. Lima, Daniel Lokshtanov, Saket Saurabh, Roohani Sharma: "Odd Cycle Transversal on P_5-free Graphs in Quasi-polynomial Time", in Proc. of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024, pp. 5276–5290, SIAM, 2024.
URL https://doi.org/10.1137/1.9781611977912.189
Main reference Daniel Lokshtanov, Paweł Rzążewski, Saket Saurabh, Roohani Sharma, Meirav Zehavi: "Maximum Partial List H-Coloring on P_5-free graphs in polynomial time", CoRR, Vol. abs/2410.21569, 2024.
URL https://doi.org/10.48550/ARXIV.2410.21569

In this talk we report a successful resolution of an open problem from the previous edition (Dagstuhl Seminar 22481) of this seminar.

At Dagstuhl Seminar 22481, Michal Pilipczuk asked whether one can find a maximum 2-colorable induced subgraph on P_5-free graphs in polynomial time or even in quasi-polynomial time.

We [Agrawal, Lima, Lokshtanov, Saurabh, Sharma] designed a quasi-polynomial time algorithm for the problem, which appeared at SODA 2024 [1]. Later together with Rzążewski, we improved the running time to polynomial for the same problem. This work will appear at ACM TALG 2025.

In a follow-up work [2], we [Lokshtanov, Rzążewski, Saurabh, Sharma, Zehavi] show that one can also design a polynomial time algorithm on P_5-free graphs for an even more general problem: the Maximum k-Colorable Induced Subgraph problem, for any fixed positive integer k. This can be further generalised to the Maximum Weight Partial List H-Coloring problem.

References

1 Akanksha Agrawal, Paloma T. Lima, Daniel Lokshtanov, Saket Saurabh, Roohani Sharma. *Odd Cycle Transversal on P_5-free Graphs in Quasi-polynomial Time*. Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms (SODA 2024), D. P. Woodruff, Ed., Alexandria, VA, USA: SIAM, 2024, pp. 5276–5290. doi:10.1137/1.9781611977912.189.
2 Daniel Lokshtanov, Paweł Rzążewski, Saket Saurabh, Roohani Sharma, Meirav Zehavi. *Maximum Partial List H-Coloring on P_5-free graphs*. CoRR, abs/2410.21569, 2024.

3.14 A survey of layered wheels

Nicolas Trotignon (CNRS – Ecole Normale Supérieure de Lyon, FR)

Layered wheels are constructions of graphs first discovered by Ni Luh Dewi Sintiari and the speaker. They disprove several conjectures and answer several questions. For instance, jointly with Maria Chudnovsky, the speaker could use them to disprove a conjecture de Dallard, Milanič and Štorgel about the tree-independence number and a conjecture of Hajebi about the induced subgraphs contained in graphs of high treewidth. The goal of the talk is to present all the situations where layered wheels turned out to be useful and to present in more detail the variant that disproved the two conjectures mentioned above.

3.15 Weighted Graph Problems and Protrusions

Michal Wlodarczyk (University of Warsaw, PL)

Main reference Michal Wlodarczyk: "Losing Treewidth In The Presence Of Weights", in Proc. of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2025, New Orleans, LA, USA, January 12-15, 2025, pp. 3743–3761, SIAM, 2025.
URL https://doi.org/10.1137/1.9781611978322.125

I will talk about vertex-deletion problems that become harder when the vertices are weighted. This is because we do not know a weighted counterpart of the technique known as "protrusion replacement". I will talk about a recent work on approximation for Weighted Treewidth-η Deletion which circumvents this issue. Then I will move on to some open problems related to weighted graphs and protrusions.

4 Open problems

4.1 Hardness/tractability radius of Hamiltonian Cycle

Édouard Bonnet (ENS – Lyon, FR)

Main reference Édouard Bonnet: "Answering Related Questions", CoRR, Vol. abs/2501.10633, 2025.
URL https://doi.org/10.48550/ARXIV.2501.10633

Given a problem Π, a distance dist over its inputs, and a function $d : N \to \mathbb{R}_+ \cup \{\infty\}$, $Sidestep(\Pi, dist, d)$ is the computational problem that takes an input I of Π, and is required to output a pair $(J, \Pi(J))$ where $dist(I, J) \leq d(|I|)$, and $\Pi(J)$ is a correct answer to Π on input J. This framework is introduced and motivated in [1].

We give one example of an open question in this setting (more are given in the preprint). $Sidestep(HamiltonianCycle, dist_e, n/3)$ is polynomial-time solvable, where $dist_e$ is the edge edit distance (i.e., the minimum number of edges to edit to go from one graph to the other) while $Sidestep(HamiltonianCycle, dist_e, n^{1/2-o(1)})$ is NP-hard. What is the "smallest" d such that $Sidestep(HamiltonianCycle, dist_e, d)$ is tractable?

References

1 Édouard Bonnet. *Answering Related Questions*. (2025), Preprint available on arxiv `https://arxiv.org/abs/2501.10633`

4.2 Polylogarithmic bounds on treewidth

Maria Chudnovsky (Princeton University, US)

For every positive integer t there exists an integer c_t with the following property. Let G be a graph with no complete subgraph of size t, and no induced minor isomorphic to $K_{t,t}$ is the txt wall. Then the treewidth of G is bounded from above by $c_t \log^{c_t} |V(G)|$.

4.3 Atoms vs. Avoiding Simplicial Vertices

Konrad Dabrowski (Newcastle University, GB)

Joint work of Konrad Dabrowski, Karl Boddy, Daniël Paulusma
Main reference Karl Boddy, Konrad K. Dabrowski, Daniël Paulusma: "Atoms Versus Avoiding Simplicial Vertices", in Proc. of the Algorithms and Complexity – 14th International Conference, CIAC 2025, Rome, Italy, June 10-12, 2025, Proceedings, Part II, Lecture Notes in Computer Science, Vol. 15680, pp. 299–315, Springer, 2025.
URL https://doi.org/10.1007/978-3-031-92935-9_19

A clique cut-set in a connected graph is a clique whose deletion results in a disconnected graph. A graph is an atom if it has no clique cut-set. A simplicial vertex is one whose neighbourhood is a clique. If an atom is not a complete graph, then it cannot contain any simplicial vertices.

For a hereditary graph class $\mathcal{G}$, we denote its set of connected graphs by $\mathcal{G}^c$, its set of atoms by $\mathcal{G}^c_a$, and its set of connected graphs that are either an atom or else have no simplicial vertices by $\mathcal{G}^c_s$. Equivalently, $\mathcal{G}^c_s$ consists of all connected graphs that are either complete or have no simplicial vertices. Note that $\mathcal{G}^c_a \subseteq \mathcal{G}^c_s \subseteq \mathcal{G}^c$. A hereditary graph class $\mathcal{G}$ will have $\mathcal{G}^c_a \neq \mathcal{G}^c_s$ if and only if G contains at least one graph from one of 27 infinite families of minimal graphs that have no simplicial vertices, but are not atoms. Roughly speaking, the graphs in these families consist of exactly two holes, which are linked together by a clique cut-set that has size at most 4 and that may contain some vertices of the two holes.

There are many computational problems that are polynomial-time solvable on a hereditary graph class $\mathcal{G}$ whenever they are polynomial-time solvable on the atoms in $\mathcal{G}^c_a$. If we replace $\mathcal{G}^c_a$ with $\mathcal{G}^c_s$ in the previous sentence, does the statement apply to a strictly wider range of computational problems?

More formally, does there exist a natural graph problem Π, for which the following holds:

1. for every hereditary graph class $\mathcal{G}$, the problem Π is polynomial-time solvable on $\mathcal{G}$ if and only if it is polynomial-time solvable on $\mathcal{G}^c_s$, and
2. there exists a hereditary graph class $\mathcal{F}$ such that Π is NP-hard on $\mathcal{F}$, but polynomial-time solvable on $\mathcal{F}^c_a$.

4.4 Counting Graphs with Vertex Cover of Size k

Daniel Lokshtanov (University of California - Santa Barbara, US)

Given as input two positive integers n and k, compute the number of graphs with vertex set $\{1, \ldots, n\}$ and at least one vertex cover of size at most k. The naive algorithm does this in time 2^{n^2}. Does there exist an algorithm with running time $2^{o(n^2)}$?

4.5 Longest Cycle Problem Parameterized Above Combinatorial Bounds

Petr A. Golovach (University of Bergen, NO)

The task of the classical Longest Cycle problem is, given a graph G and a positive integer k, to decide whether G has a cycle of length at least k. This problem generalizes Hamiltonian Cycle and is well-known to be NP-complete. From the positive side, Longest Cycle is known to be FPT when parameterized by k, and there is a long history of research focused on developing various algorithmic tools for the problem. On the other hand, various lower bounds on the circumference of a graph, i.e., the length of the longest cycle, are well-known in Extremal Combinatorics. This raises the question of whether it is possible to combine the strengths of both areas. Recently [4, 5], Longest Cycle was investigated for the parameterization above the classical circumference lower bounds of Dirac [2] and Erdős and Gallai [3], respectively. These results are obtained for undirected graphs and the area of directed graphs is completely unexplored. This leads to a plethora of open problems. For example, any directed graph of minimum in-degree $\delta^+(G) \geq 1$ has a cycle of length at least $\delta^+(G) + 1$. Is it possible to find a cycle of length $\delta^+(G) + k$ in FPT in k time on (2-strong) digraphs? More generally, we ask whether it is possible to obtain interesting and nontrivial parameterized algorithms for Longest Cycle on directed graphs above combinatorial bounds and refer to the survey of Bermond and Thomassen [1] for the circumference bounds.

References

1 Jean-Claude Bermond and Carsten Thomassen. Cycles in digraphs- a survey. *J. Graph Theory*, 5(1):1–43, 1981.
2 G. A. Dirac. Some theorems on abstract graphs. *Proc. London Math. Soc.*, (3), 2:69–81, 1952.
3 P. Erdős and T. Gallai. On maximal paths and circuits of graphs. *Acta Math. Acad. Sci. Hunger*, 10:337–356, 1959.
4 F. V. Fomin, P. A. Golovach, D. Sagunov, and K. Simonov. Algorithmic extensions of Dirac's theorem. *In Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, January 9-12, 2022*, SIAM, 2022, pp. 931–950.
5 F. V. Fomin, P. A. Golovach, D. Sagunov, and K. Simonov, Longest Cycle above Erdős-Gallai Bound. *SIAM J. Discret. Math.*, 38(4), 2721–2749, 2024.

4.6 Single-exponential time algorithms parameterized by clique-width without k-expression

Tuukka Korhonen (University of Copenhagen, DK)

When a k-expression witnessing that a graph has clique-width at most k is given, many NP-hard graph problems can be solved in time $2^{O(k)}n^{O(1)}$. Can we obtain such algorithms parameterized by clique-width without the assumption that a k-expression is given as an input? Oum, Sæther, and Vatshelle gave $2^{O(k \log k)}n^{O(1)}$ time algorithms in [1].

References

1 Sang-il Oum, Sigve Hortemo Sæther, Martin Vatshelle. *Faster Algorithms For Vertex Partitioning Problems Parameterized by Clique-width. Theoretical Computer Science*, 535:16–24, 2014. https://doi.org/10.1016/j.tcs.2014.03.024. https://arxiv.org/abs/1311.0224

4.7 The complexity of k-Induced Disjoint Paths on H_3-subgraph-free graphs

Barnaby Martin (Durham University, GB), Daniel Paulusma (Durham University, GB), Mark Siggers (Kyungpook National University – Daegu, KR), and Siani Smith (University of Bristol, GB)

Let H_i be the graph obtained from two copies of P_3 with a path of length i joining the central vertices. For k, fixed, the k-Induced Disjoint Paths problem takes as input a graph with k terminal pairs $(s_1, t_1), \ldots, (s_k, t_k)$ and asks if there are vertex-disjoint paths connecting the terminal pairs such that there are no edges between these paths.

It is known that k-Induced Disjoint Paths is in P for H_1-subgraph-free graphs and for H_2-subgraph-free graphs. Furthermore, it is NP-complete for H_i-subgraph-free graphs for all $i \geq 4$ [1]. What is the complexity in the H_3-subgraph-free case?

References

1 Vadim V. Lozin, Barnaby Martin, Sukanya Pandey, Daniël Paulusma, Mark H. Siggers, Siani Smith and Erik Jan van Leeuwen. *Complexity Framework for Forbidden Subgraphs II: Edge Subdivision and the "H"-Graphs*. 35th International Symposium on Algorithms and Computation, ISAAC 2024, December 8-11, 2024, Sydney, Australia.

4.8 Feedback vertex set (and similar problems) in P_t-free bipartite graphs

Marcin Pilipczuk (University of Warsaw, PL)

(k, ϕ)-Maximum Weight Induced Subgraph is a meta-problem that asks for a maximum-weight induced subgraph of treewidth at most k that models an CMSO2 formula ϕ; this problem captures Maximum Independent Set and Feedback Vertex Set, among others. It is known to be solvable in quasi-polynomial time in P_t-free graphs [2] and we conjecture it is actually polynomial-time solvable in these graph classes; the conjecture is confirmed up to $t = 6$ [1].

A recent manuscript [3] established this conjecture in P_7-free graphs of bounded clique number. Somewhat surprisingly, the case of P_7-free bipartite graphs turned out to be an important subcase.

This motivates the following question: can we establish polynomial-time solvability of (k, ϕ)-Maximum Weight Induced Subgraph in P_t-free bipartite graphs? A special case of Feedback Vertex Set is already very interesting. (Note that Maximum Independent Set is polynomial-time solvable in bipartite graphs thanks to the maximum matching techniques.)

References

1 Maria Chudnovsky, Rose McCarty, Marcin Pilipczuk, Michał Pilipczuk, and Paweł Rzążewski. Sparse induced subgraphs in P_6-free graphs. In David P. Woodruff, editor, *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pages 5291–5299. SIAM, 2024.

2 Peter Gartland, Daniel Lokshtanov, Marcin Pilipczuk, Michał Pilipczuk, and Paweł Rzążewski. Finding large induced sparse subgraphs in $C_{>t}$-free graphs in quasipolynomial time. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 330–341. ACM, 2021.

3 Maria Chudnovsky, Jadwiga Czyżewska, Kacper Kluk, Marcin Pilipczuk, and Paweł Rzążewski Sparse induced subgraphs in P_7-free graphs of bounded clique number. arXiv:2412.14836, 2024.

4.9 Detecting almost complete induced minors

Nicolas Trotignon (CNRS – Ecole Normale Supérieure de Lyon, FR)

For some fixed graph H, one may wonder whether detecting H as an induced minor can be done in polynomial time. For some H's, the problem is known to be polytime solvable, for others it is known to be NP-complete, so that a dichotomy theorem would be good, and this is a well-known open question that is widely open. A direction might therefore be to find graph some graphs H that are likely to give insight on the general goal. I propose $H = K_t \setminus e$, that is the graph obtained from the complete graph on t vertices by removing one edge.

These are very "close" to be polytime to detect because K_t is. Indeed, containing K_t as an induced minor is equivalent to containing it as a minor. So, to detect K_t, one may rely on the classical Robertson and Seymour algorithm to detect a minor. Having $H = K_t \setminus e$ being NP-complete would therefore be striking, but not unlikely since adding a non-edge constraint is really demanding a lot. On the other hand, it might be polytime by some clever reduction to detect a minor, or by a structure theorem for $K_t \setminus e$-induced-minor-free graphs, or by some other mean.

Participants

- Akanksha Agrawal
Indian Institute of Techology Madras, IN
- Bogdan Alecu
University of Leeds, GB
- Édouard Bonnet
ENS – Lyon, FR
- Maria Chudnovsky
Princeton University, US
- Julien Codsi
Princeton University, US
- Konrad Dabrowski
Newcastle University, GB
- Esther Galby
TU Hamburg, DE
- Jan Goedgebeur
KU Leuven, BE
- Petr A. Golovach
University of Bergen, NO
- Meike Hatzel
IBS – Daejeon, KR
- Lars Jaffke
Norwegian School of Economics – Bergen, NO
- Bart Jansen
TU Eindhoven, NL
- Jorik Jooken
KU Leuven, BE
- Tuukka Korhonen
University of Copenhagen, DK
- Stefan Kratsch
HU Berlin, DE
- Michael Lampis
University Paris-Dauphine, FR
- Paloma Lima
IT University of Copenhagen, DK
- Daniel Lokshtanov
University of California – Santa Barbara, US
- Barnaby Martin
Durham University, GB
- Martin Milanic
University of Primorska, SI
- Pranabendu Misra
Chennai Mathematical Institute, IN
- Andrea Munaro
University of Parma, IT
- Daniel Neuen
MPI für Informatik – Saarbrücken, DE
- Jelle Oostveen
Utrecht University, NL
- Sukanya Pandey
RWTH Aachen, DE
- Fahad Panolan
University of Leeds, GB
- Daniel Paulusma
Durham University, GB
- Marcin Pilipczuk
University of Warsaw, PL
- Michal Pilipczuk
University of Warsaw, PL
- Paweł Rzążewski
Warsaw University of Technology, PL
- Saket Saurabh
The Institute of Mathematical Sciences – Chennai, IN
- Oliver Schaudt
Bayer AG – Leverkusen, DE
- Roohani Sharma
MPI für Informatik – Saarbrücken, DE
- Mark Siggers
Kyungpook National University – Daegu, KR
- Siani Smith
University of Bristol, GB
- Ramanujan Sridharan
University of Warwick – Coventry, GB
- Csaba Tóth
California State University – Northridge, US
- Nicolas Trotignon
CNRS – Ecole Normale Supérieure de Lyon, FR
- Kristina Vuskovic
University of Leeds, GB
- Michal Wlodarczyk
University of Warsaw, PL
- Viktor Zamaraev
University of Liverpool, GB

Report from Dagstuhl Seminar 25042

Online Privacy: Transparency, Advertising, and Dark Patterns

Günes Acar*[1], Nataliia Bielova*[2], Zubair Shafiq*[3], and Frederik Zuiderveen Borgesius*[4]

1 **Radboud University Nijmegen, NL.** g.acar@cs.ru.nl
2 **Inria centre at University Côte d'Azur – Sophia Antipolis, FR.** nataliia.bielova@inria.fr
3 **University of California – Davis, US.** zubair@ucdavis.edu
4 **Radboud University Nijmegen, NL.** frederik.zuiderveenborgesius@ru.nl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25042 "Online Privacy: Transparency, Advertising, and Dark Patterns". The seminar brought 26 participants in computer science, law and policy together, coming from research institutions, as well as industry, law firms and regulators across Europe, US, and Middle East.

The 2.5-day seminar had a well-filled program, with introductions of all participants and several group activities; two presentations from industry representing Web browser providers, such as Apple and Mozilla; two presentations from the law research community presenting open problems in Web tracking, dark patterns, ad tech and new EU regulations, such as the EU Digital Services Act; and two panels – one presenting the open challenges in compliance by EU and US lawyers and regulators, and one discussing the future of advertising by industrial representatives from Web browser vendors. The program also included a rump session for short talks, allowing all participants to expose their recent research, open questions, and challenges to these research communities, industry, and regulators.

Seminar January 19–22, 2025 – https://www.dagstuhl.de/25042

2012 ACM Subject Classification Security and privacy → Browser security; Security and privacy → Human and societal aspects of security and privacy; Security and privacy → Privacy-preserving protocols; Security and privacy → Pseudonymity, anonymity and untraceability; Security and privacy → Social network security and privacy; Security and privacy → Web application security; Networks → Web protocol security; Information systems → World Wide Web

Keywords and phrases advertising, dark patterns, data protection, online tracking, privacy, world wide web

Digital Object Identifier 10.4230/DagRep.15.1.122

* Editor / Organizer

Online Privacy: Transparency, Advertising, and Dark Patterns, *Dagstuhl Reports*, Vol. 15, Issue 1, pp. 122–135
Editors: Günes Acar, Nataliia Bielova, Zubair Shafiq, and Frederik Zuiderveen Borgesius

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

1 Executive Summary

Günes Acar (Radboud University Nijmegen, NL)
Nataliia Bielova (Inria centre at University Côte d'Azur – Sophia Antipolis, FR)
Zubair Shafiq (University of California – Davis, US)
Frederik Zuiderveen Borgesius (Radboud University Nijmegen, NL)

The Dagstuhl Seminar on *Online Privacy: Transparency, Advertising, and Dark Patterns* enhanced the collective understanding of changes in online tracking, advertising, and dark patterns within an interdisciplinary research community in computer science and law. Following the success of the 2017 Dagstuhl Seminar "Online Privacy and Web Transparency", this seminar brought together experts from academia, legal professionals, regulators, and industry to tackle novel challenges emerging from the shifting technological and regulatory landscape.

More than half a decade after the 2017 seminar, some of the familiar questions have resurfaced in new contexts, such as smart devices and augmented reality. The introduction of privacy regulations and enforcement regimes around the world has prompted and enabled a slew of new research. Meanwhile, browsers and mobile platforms have started shipping built-in anti-tracking features, and a once-in-a-generation redesign of the online advertising and tracking ecosystem is underway.

Spanning three days, the seminar featured a variety of session formats including short talks, interactive demos, and moderated brainstorming sessions. Bringing together researchers and industry experts, the seminar promoted collaboration and advanced research on these challenges, while also exploring future research directions. Topics that were discussed during the seminar included the following:

Online Tracking Beyond Cookies

- How should online tracking research respond to fundamental changes in tracking mechanisms after the third-party cookie phaseout?
- Which techniques, tools, and methods could prove beneficial and are currently absent in researchers' toolboxes?
- Do existing regulations provide sufficient protection against novel types of tracking and profiling?
- How can computer science research help regulators with enforcement of regulations, or with developing better regulations?

Alternative Advertising Mechanisms

- What are the potential abuses associated with the novel advertising mechanisms (e.g., Topics API, Protected Audience API), and what measures can be implemented to monitor and prevent them?

Dark Patterns and Online Manipulation

- What methods and strategies are effective for detecting privacy-related dark patterns in various contexts, and how do these dark patterns influence both immediate and future privacy decisions of users?

Structure of the 3-day seminar

Day 1 – morning A plenary opening session laying the ground for the seminar, presentation of the main research topics and statistics on the participants topic interest, field of work, and background, was given to quickly foster exchanges since computer science and law experts often express the will to exchange but do not have the opportunity. Therefore, the seminar started with two sessions that enabled everybody to introduce themselves to the others. The morning session was followed by informal voting on the participants' interests in the proposed main topics and background. Topics for further discussion in working groups were identified.

Day 1 – afternoon The first session consisted of group activities on identified topics and a report of the main outcomes in the plenary session. The afternoon continued with the presentation of a browser vendor representative presenting the open problems in web tracking from an industry perspective. The last session in the afternoon featured a panel with lawyers and regulators discussing open problems in compliance, regulation, and enforcement from the EU and US perspectives.

Day 2 – morning The first session of presentations from an academic researcher on privacy signals and a browser vendor were appreciated by the participants. The morning session was followed by further group activities to discuss new topics of interest that evolved since Day 1. The morning finished with a wrapping-up session presenting the results of each group discussion.

Day 2 – afternoon The afternoon contained two sessions: presentations from legal scholars on the challenges and advancements in the EU law and new forms of transdisciplinary research between computer science and law researchers. The second session offered a panel with browser vendors, presenting unique insights into their challenges with online tracking, regulation and dark patterns to the audience.

Day 3 – morning On Day 3, there was a session with short (5 minutes) rump session talks. There was also a collective discussion on the main takeaways of the seminar with collection of feedback from the participants. It was followed by an informal session to foster further exchanges between participants who have identified common topics of interest.

Results, summary

- The seminar enhanced the collective understanding of changes in online tracking, advertising and dark patterns within an interdisciplinary research community in computer science and law.
- The seminar built a community of researchers from different disciplines who are interested in online privacy, and established further exchanges with industry and regulators.
- The seminar fostered transdisciplinary cooperation for research and future grant proposals, such as EU grants (EU collaborative projects and ERC synergy grant), bilateral agreement grants within EU countries but also EU-US, and EU-India.
- The seminar raised awareness among participating researchers about the challenges and opportunities for collaboration across computer science and law disciplines, leading to better understanding of empirical research.
- During the seminar, several participants formed interdisciplinary teams to collaborate on papers and grant proposals in the future. One of the already visible outcomes of the seminar is a new article co-authored by several participants surveying the advances and open problems in web tracking [1].

- A collaboration that started at our seminar led to the discovery of a previously undocumented tracking method, used by Meta and Yandex to track billions of Android users. The investigation led by two attendees of our seminar resulted in defenses deployed by browser vendors including Chrome and Firefox, and termination of the tracking campaign by the companies [2].

References

1 SoK: Advances and Open Problems in Web Tracking. Y. Vekaria, Y. Beugin, S. Munir, G. Acar, N. Bielova, S. Englehardt, U. Iqbal, A. Kapravelos, P. Laperdrix, N. Nikiforakis, J. Polakis,F. Roesner, Z. Shafiq, S. Zimmeck. Online report, June 2025. https://arxiv.org/abs/2506.14057

2 Covert Web-to-App Tracking via Localhost on Android. Aniketh Girish, Günes Acar, Narseo Vallina-Rodriguez, Nipuna Weerasekara, Tim Vlummens. Online report, June 2025. https://localmess.github.io

2 Table of Contents

3 Overview of Talks

3.1 Dark Patterns as Legal Violations in Web Tracking

Cristiana Santos (Utrecht University, NL)

Joint work of Cristiana Santos, Nataliia Bielova, Colin Gray, Johana Gunawan, Sanju Ahuja, Christine Utz
Main reference Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, Thomas Mildner: "An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building", in Proc. of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024, pp. 289:1–289:22, ACM, 2024.
URL https://doi.org/10.1145/3613904.3642436

Extant regulations worldwide govern dark patterns explicitly or implicitly. EU member states and US states' own statutes and enforcers, as well as other union-wide legislation or authorities may also regulate some types of dark patterns or related behaviours – or otherwise issue guidance. A growing body of enforcement actions and regulatory fines globally currently comprise a strong approach for dark patterns general deterrence [1]. In this talk, I discuss web tracking practices that may constitute legal violations under the GDPR, ePD, and can be aligned with dark patterns. Aligning non-compliant online tracking practices with dark pattern prohibitions enhances dark pattern general deterrence. I report several instances thereof based on our legal-empirical research.

- Publishers and CMPs don't respect users' choice [2, 3]:
  - consent banner stores a positive consent even when the user refused consent, corresponding to the dark pattern of sneaking;
  - a positive consent is stored before the user made a choice, corresponding to the dark pattern of sneaking;
  - the consent request does not offer a way to refuse consent, corresponding to the dark pattern of obstruction;
  - Some purposes or advertisers are pre-selected: pre-ticked boxes or sliders set to "accept", corresponding to the dark pattern of bad defaults.
- CMP website scanners are used as compliance solutions, though these introduce:
  - false negatives: only scans cookies, but miss other tracking technologies, such as browser fingerprinting, and as such, data is processed without legal basis [4], which corresponds to the dark pattern of 'hidden information';
  - false positives: scanners deceive editors that a consent banner is needed on an empty website without any trackers [5], aligned with the dark pattern of 'forced action'.
- The QuantCast CMP banner sets and sends QuantCast cookie to its server without a legal basis, potentially infringing the lawfulness and fairness principles, and this practice can qualify the dark pattern of sneaking and forced action [4].
- Pay or ok models offer 2 options to end-users in order to gain access to an online service: i) consent to being tracked and targeted with behavioural advertising, or ii) pay a ad-tracking fee. Dark Patterns also occur in the "Pay or Ok models" [6] where social engineering dark patterns appear under the pay option.
- Google Tag Manager (GTM) facilitates inclusion of third-party JS and is used on 62% on top of 100k websites. Within the GTM ecosystem, 780 not-supported Google tags are hidden, Google-owned tags are instead featured on top, and 67 tags supported by Google are only shown at the bottom, which configure the dark patterns of false-hierarchy and Adding Steps [7].

References

1 An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building, 2024. Colin M. Gray, Cristiana Santos, Nataliia Bielova, and Thomas Mildner. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24).
2 Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework, 2020. Célestin Matte, Nataliia Bielova, Cristiana Santos. IEEE Symposium on Security and Privacy, 2020.
3 Dark Patterns and the Legal Requirements of Consent Banners: An Interaction Criticism Perspective, 2021. Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, Damian Clifford. ACM CHI Conference on Human Factors in Computing Systems (CHI '21).
4 Consent Management Platforms under the GDPR: processors and/or controllers? 2021. Cristiana Santos, Midas Nouwens, Michael Toth, Nataliia Bielova, Vincent Roca. In Privacy Technologies and Policy: 9th Annual Privacy Forum.
5 On dark patterns and manipulation of website publishers by CMPs. Michael Toth, Nataliia Bielova, Vincent Roca. Privacy Enhancing Technologies Symposium, 2022.
6 Legitimate Interest is the New Consent – Large-Scale Measurement and Legal Compliance of IAB Europe TCF Paywalls. Victor Morel, Cristiana Santos, Viktor Fredholm, and Adam Thunberg. 2023. In Proceedings of the 22nd Workshop on Privacy in the Electronic Society (WPES '23).
7 Which Online Platforms and Dark Patterns Should Be Regulated under Article 25 of the DSA? 2024, Cristiana Santos, Nataliia Bielova, Sanju Ahuja, Christine. Utz, Colin Gray, Gilles Mertens, Preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4899559

3.2 Studying Privacy Threats in Complex and Interconnected Platforms: The Case of Smart Homes

Narseo Vallina-Rodriguez (IMDEA Networks Institute – Madrid, ES)

Joint work of Aniketh Girish, Tianrui Hu, Vijay Prakash, Daniel J. Dubois, Srdjan Matic, Danny Yuxing Huang, Serge Egelman, Joel Reardon, Juan Tapiador, David R. Choffnes, Narseo Vallina-Rodriguez
Main reference Aniketh Girish, Tianrui Hu, Vijay Prakash, Daniel J. Dubois, Srdjan Matic, Danny Yuxing Huang, Serge Egelman, Joel Reardon, Juan Tapiador, David R. Choffnes, Narseo Vallina-Rodriguez: "In the Room Where It Happens: Characterizing Local Communication and Threats in Smart Homes", in Proc. of the 2023 ACM on Internet Measurement Conference, IMC 2023, Montreal, QC, Canada, October 24-26, 2023, pp. 437–456, ACM, 2023.
URL https://doi.org/10.1145/3618257.3624830

Privacy risks may occur when platforms, software, and devices interact with other colluding elements in the local network using wireless interfaces like WiFi or Bluetooth[1]. However, the research community has typically followed a process-centric and monolithic approach to detect such abuses in modern consumer-oriented software, while current privacy controls are not fit to limit side-channels and covert-channels that exist in such interconnected environments.

The network communication between Internet of Things (IoT) devices on the same local network has significant implications for platform and device interoperability, security, privacy, and correctness. Yet, the privacy issues of local home Wi-Fi network traffic and its associated security and privacy threats have been largely ignored by prior literature. In this talk, I presented the results of a comprehensive and empirical measurement of the interactions that occur between devices on the local network and its threats [2]. Our analysis reveals vulnerable devices, insecure use of network protocols, and sensitive data exposure by IoT

devices. We provide evidence of how this information is exfiltrated to remote servers by mobile apps and third-party SDKs, potentially for household fingerprinting, surveillance, and cross-device tracking.

References

1 A. Girish, J. Reardon, S. Matic, J. Tapiador, and N. Vallina-Rodriguez. Your Signal, Their Data: An Empirical Privacy Analysis of Wireless-scanning SDKs in Android. PETS Symposium 2025 (To Appear)
2 A. Girish, T. Hu, V. Prakash, D. Dubois, S. Matic, D. Huang, S. Egelman, J. Reardon, J. Tapiador, D. Choffnes, and N. Vallina-Rodriguez. In the Room Where It Happens: Characterizing Local Communication and Threats in Smart Homes. ACM IMC 2023.

3.3 Global Privacy Control in EU Data Protection Laws

Sebastian Zimmeck (Wesleyan University – Middletown, US)

Joint work of Katherine Hausladen, Oliver Wang, Sophie Eng, Jocelyn Wang, Francisca Wijaya, Matthew May, Sebastian Zimmeck
Main reference Katherine Hausladen, Oliver Wang, Sophie Eng, Jocelyn Wang, Francisca Wijaya, Matthew May, Sebastian Zimmeck: "Websites' Global Privacy Control Compliance at Scale and over Time" 34th USENIX Security Symposium (USENIX Security) Seattle, CA, August 2025
URL https://sebastianzimmeck.de/hausladenEtAlGPCWeb2025.pdf

The California Consumer Privacy Act (CCPA) gives California residents the right to opt out of the sale or sharing of their personal information via Global Privacy Control (GPC). Similar other states in the US also give their residents a right to opt out via GPC. However, how can GPC be applied in the European Union? GPC is adaptable to various laws as it is does not prescribe a particular meaning to what a GPC signal means beyond the general meaning of opting out. Thus, it is the task of legislators and regulators in every jurisdiction to fill GPC with life. This talk highlighted how GPC works, its adoption in the US, and how it could work in the EU. To map GPC to the GDPR the legal basis for processing can be taken into account: (1) where the legal basis is consent, the data subject is withdrawing their consent under Article 7(3) specifically to processing by data controllers other than the first party and to processing of the first party to transfer data to other data controllers, (2) where the legal basis is legitimate interest or public interest, the data subject is objecting to processing by data controllers other than the first party and to processing of the first party to transfer data to other data controllers under Article 21(1-3, 5), and (3) where the legal basis is contractual, legal obligation, or vital interests the signal has no effect [1].

References

1 Berjon Robin. *GPC under the GDPR*. https://berjon.com/gpc-under-the-gdpr/, 2021

3.4 The EU Digital Services Act: what does it mean for online advertising and adtech?

Frederik Zuiderveen Borgesius (Radboud University Nijmegen, NL)

What does the Digital Services Act (DSA) mean for online advertising? We describe and analyse the DSA rules that are most relevant for online advertising and adtech (advertising technology). We also highlight to what extent the DSA's advertising rules add something to the rules in the General Data Protection Regulation (GDPR) and the ePrivacy Directive. The DSA introduces several specific requirements for online advertising. First, the DSA imposes transparency requirements in relation to advertisements. Second, very large online platforms (VLOPs) should develop a publicly available repository with information about the ads they presented. Third, the DSA bans profiling-based advertising (behavioural advertising) if it uses sensitive data or if it targets children.

Besides these specific provisions, the general rules of the DSA on illegal content also apply to advertising. Advertisements are a form of information, and thus subject to the general DSA rules. Moreover, we conclude that the DSA applies to some types of ad tech companies. For example, ad networks, companies that connect advertisers to publishers of apps and websites, should be considered platforms. Some ad networks may even qualify as VLOPs.

Hence, ad networks must comply with the more general obligations in the DSA. The application of these general rules to advertisements and ad networks can have far-reaching effects that have been underexplored and deserve further research. We also show that certain aspects of the DSA are still unclear. For instance, we encourage the European Commission or regulators to clarify the concepts of 'online platform' and 'recipients' in the context of ad networks and other adtech companies.

4 Panel discussions

4.1 Panel: collective discussions on main takeaways of the seminar

Günes Acar (Radboud University Nijmegen, NL), Nataliia Bielova (Inria centre at University Côte d'Azur – Sophia Antipolis, FR), Zubair Shafiq (University of California – Davis, US), and Frederik Zuiderveen Borgesius (Radboud University Nijmegen, NL)

Browser Ecosystem and Collaboration

Browsers have undergone significant evolution over time, and this progress underscores the fact that laws – and the ways in which they are interpreted – also continuously change. Engaging with actors outside of academia, such as reporters, is important for effectively disseminating research findings to broader audiences. The browser vendor panel proved to be particularly valuable, as it revealed relationships among vendors that were previously unknown. This highlights the importance of fostering collaboration and dialogue between the research community and the browser ecosystem.

Security and Standardization

There is much to be learned from the security community, particularly in how it clearly defines and responds to various behaviors. It is important to clarify which behaviors are considered good, bad, or deceptive, and to establish mechanisms to block, prevent, or mitigate the harmful ones. A key challenge lies in identifying an appropriate venue for standardizing these efforts, with peer-reviewed journals suggested as a possible avenue.

Privacy, Manipulation, and Public Concerns

Concerns have been raised about the increasing reliance on tools like ChatGPT, particularly regarding the potential for manipulation and the implications for monopolistic control. This situation emphasizes the importance of being able to articulate privacy risks with clarity, both for public understanding and for informing policy and technical responses.

Research Directions in Privacy

There is a need for further research aimed at exposing and explaining privacy risks in digital environments. One important area of study involves demonstrating how tracking can be unavoidable under current conditions. This seminar was especially convincing in showing that the study of dark patterns can be approached as a scientific discipline, worthy of systematic investigation.

Mitigation Strategies and Standards

Conducting breakage analysis could be a valuable tool in the design of effective tracking mitigation strategies. Additionally, there may be a need to develop a framework similar to the "Better Ads Standard" to guide acceptable tracking practices and support a more privacy-respecting web ecosystem.

Privacy Signals

Privacy standards are challenging to develop because individuals and organizations operate with different threat models. The talk on privacy signals, such as Global Privacy Control (GPC), was found to be particularly interesting. It raised important questions about what changes might be needed from browser vendors to better support such signals. Topics discussed included the role of dark patterns in consent dialogs, the ongoing risk posed by fingerprinting techniques, and the possibility of a class action emerging in Europe related to these privacy concerns.

Impact on regulatory compliance

Participants discussed the limitations of regulators, particularly their geographical constraints, and debated whether naming and shaming actually leads to reform or merely pushes dark patterns elsewhere. It was noted that vague warnings like "thousands of websites tracking you" are ineffective; instead, there is a need to clearly articulate specific harms and improve communication strategies. The group considered major challenges in the field, balancing pessimism with activism, and noting that even large fines, like Meta's 13 billion USD, might indicate that regulation can have some impact. There was optimism in seeing regulators' receptiveness and recognition that most users cannot be expected to understand how the underlying technology works.

Future challenges

The seminar sparked many new research directions and provided valuable insights, including developer perspectives. One attendee called it "probably the best Dagstuhl I've been to." There was reflection on whether the privacy battle has already been lost due to the vast trails of data left behind, raising the question of whether current efforts are more for the benefit of future generations. Collaborative work with browser vendors was highlighted as a path to meaningful change. Cross-disciplinary exchange – especially with legal experts – was seen as a major strength of the seminar, though it was also noted that economists were absent from the discussion. Finally, ethical questions about the strategy of naming and shaming were raised, signaling a need for deeper consideration of advocacy tactics.

4.2 Panel: Compliance, Regulation and Enforcement

Günes Acar (Radboud University Nijmegen, NL), Nataliia Bielova (Inria centre at University Côte d'Azur – Sophia Antipolis, FR), Zubair Shafiq (University of California – Davis, US), and Frederik Zuiderveen Borgesius (Radboud University Nijmegen, NL)

The panel took place at the end of Day 1 and the panelists were: Jason "Jay" Barnes (Simmons Hanly Conroy); Lesley E. Weaver (Bleichmar Fonti & Auld); Vincent Toubiana (CNIL); Frederik Zuiderveen Borgesius (moderator):

Jason "Jay" Barnes Attorney Jason "Jay" Barnes is a partner at Simmons Hanly Conroy in the Complex Litigation Department where he focuses his practice on consumer class action lawsuits. Before joining the firm, Jay served eight years as a state representative in the Missouri General Assembly. In this role, he fought against fraud, abuse and waste as chairman of the House Committee on Government Oversight and Accountability. He also served as chairman of the Special Investigative Committee on Oversight formed in 2018 to investigate the wrongdoings of former Missouri governor Eric Greitens. https://www.simmonsfirm.com/about-us/our-attorneys/jason-barnes/

Lesley E. Weaver Lesley joined Bleichmar Fonti & Auld LLP as a partner in 2016, opening the firm's California office. In her twenty year career, Lesley has focused primarily on cases that protect the public interest, consumers, and public entities. As part of her mission to protect the public trust, Lesley also serves as counsel to a number of governmental entities, in both formal and informal roles. Lesley represents the Cities of Palo Alto and Richmond, California in a municipal subclass in In re Lithium Ion Batteries Antitrust Litig. Lesley also represents Oakland County, Michigan in In re Liquid Aluminum Sulfate Antitrust Litig. Lesley is committed to public service through volunteer efforts, and currently serves on the Advisory Council of the East Bay Community Law Center, as well as the Executive Committee of the Securities Section for the Bar Association of San Francisco. https://www.bfalaw.com/professionals/lesley-weaver

Vincent Toubiana Vincent works at the CNIL, the Commission Nationale de l'Informatique et des Libertés. The CNIL is the French Data Protection Authority. Vincent has been head of CNIL's digital innovation lab (the LINC) since 2021. He obtained a PhD "Computer Science and Networks" from Telecom ParisTech in 2008. He has worked on privacy since 2009. First at NYU under the direction of Helen Nissenbaum, then from 2010 to 2013, at Alcatel-Lucent Bell-Labs. He joined CNIL in 2013 as a technologist. In 2016, he was an International fellow at the Federal Trade Commission.

Frederik Zuiderveen Borgesius Frederik is professor of ICT and law. He works at the iHub, part of Radboud University in The Netherlands. The iHub is the interdisciplinary research hub on digitalization and society. Frederik is a law professor but teaches mostly at the computer science department. His research predominantly concerns fundamental rights, such as the right to privacy and non-discrimination rights, in the context of new technologies. He often enriches legal research with insights from other disciplines. He has cooperated with, for instance, economists, computer scientists, and communication scholars. He regularly advises policymakers, and has given expert testimony at the Dutch and the European parliaments, and committees of the Council of Europe and the United Nations. `https://www.ru.nl/personen/zuiderveen-borgesius-f/`

Overview of the discussion

The panel focused on the discussion of law and legal compliance, in particular relevant law in the US and the EU. The discussed topics included lessons learned from the 5+ years of GDPR; obstacles to enforcement and litigation in the EU and US; experiences from the US case law; insights on improving the exchanges and relations between regulators and researchers; recent new regulations in the EU and the United States, such as the EU Digital Service Act (DSA), the California Privacy Rights Act (CPRA) and decisions by the US Federal Trade Commission (FTC).

Differences in the EU and US laws related to privacy

During the panel there was quite some discussion about the differences between the law in the EU and the US. For instance, in the US, (private law) court cases between groups of claimants against companies play a large role in privacy law. In the EU, such cases are rare. Meanwhile, in the EU, there are many cases in which Data Protection Authorities enforce the GDPR. Such enforcement actions sometimes lead to (administrative law) court cases between the Data Protection Authority and the company.

Exposing research results to regulators

All participants expressed interest in new empirical findings about online tracking by academic researchers. Yet, lawyers and regulators rarely have time to read through the full academic publications, pointing out that blog posts about academic findings would be more appreciated. Additionally, similarly to the FTC's PrivacyCon in-house conference, the CNIL organizes a yearly event, called Privacy Research Day, where researchers are invited to submit their contributions to achieve a higher visibility and impact of their research results.

4.3 Panel: Browser vendors: Future of tracking and advertising

Nataliia Bielova (Inria centre at University Côte d'Azur – Sophia Antipolis, FR), Günes Acar (Radboud University Nijmegen, NL), Zubair Shafiq (University of California – Davis, US), and Frederik Zuiderveen Borgesius (Radboud University Nijmegen, NL)

The panel took place at the second day of the seminar and the panelists were as follows: Igor Bilogrevic (Google); Hamed Haddadi (Imperial College London/Brave); Anastasia Shuba (DuckDuckGo); John Wilander (Apple); Martin Thomson (Mozilla), moderator.

With participants from five different browser vendors, the panel focused on the future of tracking and advertising, in light of recent regulatory changes, efforts such as Privacy Sandbox and rise of LLMs.

Participants highlighted breakage (e.g. bug due to tracking mitigations) being a significant challenge for shipping tracking defenses. It was noted that vendors could better communicate both among themselves and with wider research community.

The impact of AI taking over web search was discussed, with questions raised about the sustainability of the web in such a future. Concerns about the future of tracking (e.g. in LLM-based chat interfaces) and privacy problems that might arise ten years from now were also discussed in hypothetical terms.

Overall, the panel provided rare insights into challenges faced by the browser vendors who aim to develop and ship more tracking defenses. In the post-seminar survey, several participants indicated this panel to be one of their favorite throughout the seminar.

Participants

- Günes Acar
Radboud University
Nijmegen, NL
- Jason "Jay" Barnes
Simmons Hanly Conroy –
New York, US
- Nataliia Bielova
Inria centre at University Côte
d'Azur – Sophia Antipolis, FR
- Igor Bilogrevic
Google – Zürich, CH
- Yana Dimova
DistriNet, KU Leuven, BE
- Serge Egelman
ICSI – Berkeley, US
- Imane Fouad
INRIA Lille, FR
- Colin M. Gray
Indiana University –
Bloomington, US
- Johanna Gunawan
Maastricht University, NL
- Hamed Haddadi
Imperial College London, GB
- Martin Johns
TU Braunschweig, DE
- Konrad Kollnig
Maastricht University, NL
- Athina Markopoulou
University of California –
Irvine, US
- Rishab Nithyanand
University of Iowa –
Iowa City, US
- Cristiana Santos
Utrecht University, NL
- Zubair Shafiq
University of California –
Davis, US
- Anastasia Shuba
DuckDuckGo – Paoli, US
- Sandra Siby
New York University –
Abu Dhabi, AE
- Martin Thomson
Mozilla – Mountain View, US
- Vincent Toubiana
CNIL – Paris, FR
- Christine Utz
Radboud University
Nijmegen, NL
- Narseo Vallina-Rodriguez
IMDEA Networks Institute –
Madrid, ES
- Lesley E. Weaver
Bleichmar Fonti & Auld –
Oakland, US
- John Wilander
Apple – Cupertino, US
- Sebastian Zimmeck
Wesleyan University –
Middletown, US
- Frederik Zuiderveen Borgesius
Radboud University
Nijmegen, NL

Report from Dagstuhl Seminar 25051

Trust and Accountability in Knowledge Graph-Based AI for Self Determination

John Domingue[*1], Luis-Daniel Ibáñez[*2], Sabrina Kirrane[*3], Maria-Esther Vidal[*4], and Philipp D. Rohde[†5]

1 **The Open University – Milton Keynes, GB.** john.domingue@open.ac.uk
2 **University of Southampton, GB.** L.D.Ibanez@soton.ac.uk
3 **Vienna University of Economics and Business, AT.** sabrina.kirrane@wu.ac.at
4 **TIB – Hannover, DE.** vidal@l3s.de
5 **TIB – Hannover, DE.** philipp.rohde@tib.eu

Abstract

This report documents the program and results of the Dagstuhl Seminar 25051 "Trust and Accountability in Knowledge Graph-Based AI for Self Determination". The seminar focused on AI systems powered by Knowledge Graphs and their fundamental role in powering intelligent decision making. Knowledge Graphs complement Machine Learning algorithms by providing data context and semantics, enabling further inference and question answering capabilities, and their synergy with Large Language Models is being actively researched. Despite the numerous benefits that can be accomplished with KG-based AI, its growing ubiquity within online services may raise the loss of self-determination for citizens as a fundamental societal issue. The more we rely on these technologies, which are often centralised, the less citizens will be able to determine their own destiny. To counter this threat, AI regulation, such as the EU AI Act, is being proposed in certain regions. Regulation sets what technologists need to do, leading to questions concerning: How can the output of AI systems be trusted? What is needed to ensure that the data fueling and the inner workings of these artefacts are transparent? How can AI be made accountable for its decision-making?

Seminar January 26–31, 2025 – https://www.dagstuhl.de/25051

2012 ACM Subject Classification Information systems → Decision support systems; Theory of computation → Semantics and reasoning; Information systems → Graph-based database models; Computing methodologies → Artificial intelligence

Keywords and phrases access control and privacy, federated query processing, intelligent knowledge graph management, programming paradigms for knowledge graphs, semantic data integration

Digital Object Identifier 10.4230/DagRep.15.1.136

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

Trust and Accountability in Knowledge Graph-Based AI for Self Determination, *Dagstuhl Reports*, Vol. 15, Issue 1, pp. 136–200
Editors: John Domingue, Luis-Daniel Ibáñez, Sabrina Kirrane, Maria-Esther Vidal, and Philipp D. Rohde

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

1 Executive Summary

John Domingue (The Open University – Milton Keynes, GB, john.domingue@open.ac.uk)
Luis-Daniel Ibáñez (University of Southampton, GB, L.D.Ibanez@soton.ac.uk)
Sabrina Kirrane (Vienna University of Economics and Business, AT, sabrina.kirrane@wu.ac.at)
Maria-Esther Vidal (TIB – Hannover, DE, vidal@l3s.de)

In just one minute in April 2022, there were 5,900,000 searches on Google, 1,700,000 pieces of content shared on Facebook, 1,000,000 hours streamed and 347,200 tweets shared on Twitter.[1] This content and data is linked to a plethora of AI services which have increasingly been based on Knowledge Graphs (KGs), i.e., machine-readable data and schema representations based on a web stack of standards. The term "Knowledge Graph" was first introduced by Google in 2012 and is strongly linked to the work of the Semantic Web community which first began in around 2001 based on the seminal paper Berners-Lee et al.[2]. The main types of areas covered by AI services include, for example, content recommendation, user input prediction, and large-scale search and discovery and form the basis for the business models of companies like Google, Netflix, Spotify, and Facebook.

Over a number of years, there has been a growing worry on how personal data can be abused and thus, how AI services impinge on citizens' rights. For example, the over centralisation of data and linked abuses led Sir Tim Berners-Lee to call the Web "anti-human" in an interview in 2018[3] and since 2016, hundreds of US Immigration and Customs Enforcement employees have faced investigations into abuse of confidential law enforcement databases including stalking and harassment, up to passing data to criminals.[4]

The subject of proposed legislation today is ensuring that digital platforms, including AI platforms, provide societal benefit. Within Europe, the proposed EU AI Act aims to support safe AI that respects fundamental human rights. The regulation sets what technologists need to do. Our seminar was structured around three pillar research topics – trust, accountability, and self-determination – that represent the desired goals, and four foundational research topics – Machine-readable Norms and Policies, Decentralised KG Management, Neuro-Symbolic AI, and Decentralised Applications – that constitute the necessary technical foundations to achieve the goals.

[1] https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/
[2] Berners-Lee, T., Hendler, J. and Lassila, O., The semantic web. Scientific American, 284(5), pp. 34-43.
[3] "I Was Devastated": Tim Berners-Lee, the Man Who Created the World Wide Web, Has Some Regrets | Vanity Fair
[4] https://www.wired.com/story/ice-agent-database-abuse-records/

2 Table of Contents

3 Invited Talks

3.1 Why are we still not there (The Semantic Web today)

James A. Hendler (Rensselaer Polytechnic Institute – Troy, US)

In 2001, Tim Berners-Lee, Ora Lassila and I wrote a vision paper in Scientific American outlining potential challenges on the Web and exploring how ontologies and (what are now called) knowledge graphs could improve interoperability among Web systems and allow agents on the Web to cooperatively solve problems. The paper opened with a scenario of two people coordinating their parent's medical treatment and involved a number of different web sites, but a relatively straightforward task. In this talk, I reminded people of that challenge, and pointed out that despite all that has happened in AI, in Semantic Web, and in "agentic systems", we still cannot do the simple task that was outlined in that paper nearly 25 years ago. This talk discussed why, and what we might think about going forward to solve such problems. The Semantic Web paper grew in part out of a 2000 Dagstuhl Seminar (00121) and led to a number of later meetings that have continued on the theme. The 2001 paper is the most academically cited paper in the history of the Scientific American publication, and a number of Dagstuhl meetings in the years since have focused on Semantic Web and related topics. Some of the most cited Dagstuhl Seminars have appeared in this series (such as 24061) and this most recent seminar was a wonderful opportunity to continue the series – so this talk also acknowledged the huge role Dagstuhl has had in creating and sustaining this community over the past 25 years.

4 4-Minute Talks

4.1 Neuro-symbolic, agentic AI for organizing scholarly knowledge

Sören Auer (TIB – Hannover, DE)

The exponential growth of scientific publications poses a significant challenge for researchers, policymakers, and automated systems alike: how to effectively access, structure, and reason over ever-expanding bodies of knowledge. We aim to leverage neuro-symbolic and agentic AI to transform the organization and consumption of scholarly knowledge. At the core of this vision is the Open Research Knowledge Graph (ORKG) – an infrastructure designed to capture, interlink, and semantically represent the core contributions of scientific articles. Unlike traditional bibliographic databases, the ORKG enables structured comparisons of research contributions, supports semantic search, and integrates knowledge across disciplines. Building on this foundation, ORKG ASK provides an intelligent assistant that can answer complex research questions by composing and contextualizing information across the graph, significantly enhancing discoverability and scientific insight. We explore how neuro-symbolic methods – combining deep learning with semantic representations over knowledge graphs – can be further augmented by agentic AI. These agents autonomously navigate, enrich, and reconcile scholarly knowledge by performing tasks such as hypothesis mapping, claim validation, and comparison synthesis. This synergy enables a new generation of interactive and explainable AI systems that actively support scientific discovery and meta-research.

4.2 Pillar #1: Meanings of Trust

Piero A. Bonatti (University of Naples, IT)

Trust in Knowledge Graphs needs to be supported with a wide range of methodologies and technologies that address complementary issues. First, KG are incresingly being used to encode confidential information and personal data that shall be appropriately protected; more generally, the use of such data shall be restricted, even after data disclosure. Second, KG contents – that is provided by manifold "agents" (both humans and AI) using diverse knowledge sources – should be reliable, and their integrity should be protected. The inferences that can be drawn from KG should be reliable as well. And the people in charge of maintaining the KG with all of its sensitive information should be trustable, too.

A closer look at the above points reveals several connections with the other pillars mentioned in this seminar's manifesto, for example:

- Accountability, like trust in KG content and KG management, may leverage logging, provenance, integrity preserving and non-repudiation techniques.
- Self determination involves – among other aspects – the protection of personal data and control on its usage.

Some of the techniques that may help in improving trust in KG, along the above lines, are well-established (such as those for integrity preservation and non-repudiation). Some challenges (such as making AI more reliable and explainable to trust its inferences, or enforcing usage restrictions after data disclosure) are probably not specific to KG. Thus, the list of strictly trust-related challenges is not obvious and needs to be carefully drawn.

For sure, access control to, and the anonymization of, KG are harder than their traditional counterparts for relational databases, both because there is no reference schema to rely upon, and because the many inference tools that operate on KG may reveal concealed confidential data. Accordingly, the intrinsic computational complexity of access control and anonymization is often harder than in the classic case, which poses an obvious challenge.

The research on access control and anonymization techniques for KG and knowledge bases does not always take into account the large body of experience developed in the area of computer security and privacy [1]. This is a major risk that calls for a more extensive exploration of security and privacy papers. Moreover, some machine-readable policy languages that are becoming increasingly popular in the KG world, lack formal semantics. Consequently, they are ambiguous and under-specified in several respects. This is likely to jeopardize trust in many natural distributed scenarios, where different parties and stakeholders should understand policies in the same way, in order to avoid violations and sanctions.

References

1 Piero A Bonatti. A false sense of security. *Artificial Intelligence*, 310:103741, 2022.

4.3 Challenging Scenarios for Trust and Accountability in KG-based AI

Irene Celino (CEFRIEL – Milan, IT)

While Knowledge Graphs and other AI technologies provide an irrefutable advantage in managing data and knowledge in a meaningful way, several application scenarios present difficult challenges in relation to the management of trust and accountability between people and between organizations. In my short talk, I presented three different scenarios coming from my experience with their specific issues.

In the context of *cognitive-behavioural therapy*, psychological/psychiatric patients compile the so-called cognitive diaries, i.e. stories about personal events, aimed to help them reflect on emotions, thoughts, feelings and behaviours. Those diaries are shared with therapists, to enable early identification of signals that could lead to psychotic episodes. A digital version of cognitive diaries can support their processing to help therapist to promptly intervene. In my experience of digitisation of cognitive diaries[5], several issues emerged:

(1) Do patients trust a system to collect their diaries? Who has the legal right to access them (e.g. legal guardians)?

(2) Cognitive diaries may contain mentions of real/imagined events involving real/imaginary people: how to deal with those reported "facts"? They are not the same as misinformation?

(3) How to anonymise diary content (mostly textual) in a responsible way? How to share anonymised content with the scientific community without adding specific context (which may be required to correctly understand)?

In the context of *industrial procedures*, I'm currently working[6] on enabling the construction of procedural knowledge graphs, to collect employees' knowledge about industrial procedures (i.e. how-to), often based on experience, possibly including tacit knowledge. The goal is to provide the industry workers with KG-powered AI tools [1] to support their compliance with the procedures and to reduce the possible mistakes. In this case, emerging challenges are:

(1) Are workers willing to share their procedural knowledge? Do they fear losing their professional value? Do they fear being monitored?

(2) Do AI tools always ensure accuracy and reliability when providing information to industry workers? Are they safe?

(3) Are industrial employees scared of being replaced by AI? What kind of human knowledge and abilities should be preserved as such (e.g. critical thinking)?

In the context of *mobility data spaces*[7], an ecosystem of mobility actors (service providers, authorities, etc.) may be willing to cooperate (e.g. Mobility-as-a-Service), but in the meantime they are in competition and they want to preserve their competitive advantage. Therefore, they need to regulate data and service sharing in a distributed, federated (e.g. European National Access Points) and possibly "untrusted" business environment. Apart from the research challenges specifically related to the data space topic, other issue may emerge: (1) What if there are multiple, incompatible data spaces? How can a mobility actor avoid duplication of work to connect to different digital ecosystems (especially when they

[5] DIPPS project, co-funded by the Italian Ministry of Enterprises and Made in Italy
[6] PERKS project, co-funded under the EU Horizon Europe Programme (`https://perks-project.eu/`)
[7] deployEMDS project, co-funded under the EU Digital Europe Programme (`https://deployemds.eu/`)

already entered some)? (2) Are mobility actors willing to share business-critical information (even if required by laws)? Can they preserve their right to retain their information and beliefs when entering a negotiation (e.g. non disclosure of disagreement, difference between ontological agreement and ontological commitment [2])?

References

1 Irene Celino, Valentina A. Carriero, Andrea Azzini, Ilaria Baroni, and Matteo Scrocca. Procedural knowledge management in industry 5.0: Challenges and opportunities for knowledge graphs. *Journal of Web Semantics*, 84:100850, 2025.
2 Emanuele Della Valle, Irene Celino, and Davide Cerizza. Agreeing while disagreeing, a best practice for business ontology development. In *Proceedings of the 11th International Conference on Business Information Systems*. Springer, 2008.

4.4 Policies in decentralised ecosystems and academic regulatory frameworks

Andrea Cimmino (Polytechnic University of Madrid, ES)

Knowledge graphs and decentralised data has become one of the pillars of the European data-driven infrastructures like data spaces or proposals to foster data sovereignty like SOLID Pods. In this context, specifying the circumstances under which data should be accessed has become a crucial task. On the one hand, one challenge is specifying in an unambiguous way the terms, conditions, and actions that constitute a policy under which such data can be exploited in a machine-readable format. On the other hand, evaluating and enforcing policies to check whether the usage of such data legit. In this context, my personal research interests go in two lines. The first, related to the former challenge, is to express different academic regulatory frameworks as policies or norms to analyse their interoperability, quality, or the conformance of physical lessons to such regulations. The second, related to the latter challenge, is to tackle challenges derived from the evaluation and enforcement of policies in a decentralised ecosystem. In addition, it is worth researching the usage of LLMs in the different steps of the life-cycle of policy management.

4.5 Improving trustworthiness of ML through evaluation and formal guarantees

Michael Cochez (VU Amsterdam, NL)

My research focuses on neuro-symbolic AI, particularly with knowledge graphs (KG). I investigate how graph neural networks can bridge the gap between discrete KGs and the statistical world of machine learning, enabling more robust and scalable solutions for tasks like structured and natural language question answering. My work often utilizes data from the medical and biomedical domains.

During the seminar, I hope to discuss two key challenges:

Improving ML Trustworthiness: In my view current evaluation metrics are inadequate, leading to overconfidence in models. Among possible solutions, I want to discuss alternative evaluation methods and the need for formal guarantees (e.g., error bounds) to enhance trust in graph ML models.

Addressing Bias in Multi-source KGs: I will discuss the challenges of learning on unbalanced KGs, where a larger, potentially biased graph overshadows smaller, more specific ones. This hinders the use of graph ML in decentralized settings where self-determination is crucial.

4.6 Decentralisation the key for Ensuring Trust, Privacy and Personalisation in KG-based AI systems

John Domingue (The Open University – Milton Keynes, GB)

The explosion in interest and take-up of AI based systems has grown enormously since the arrival of GenAI through ChatGPT at the end of 2022. At the Open University we have been exploring how machine learning and AI can aid our 200K+ students since 2011 when we created OU Analyse[8] – a learning analytics tool able to predict if a student is at risk of failing the next assignment or course overall with 95% accuracy in the best cases. In 2015 we began investigating how decentralising technologies, such as distributed ledgers and personal data stores, such as Solid, could enable students to be "Self Sovereign" with respect to their credentials[9]. Since the beginning of 2023 we have been developing and evaluating two GenAI based tools to support teaching and learning at the OU. Our AI Module Writing Assistant (AIMWA) [1] supports OU academics in the writing of new courses. A pilot is currently underway with a module writing team in the Faculty of Business and Law who are creating a new MBA course. The AI Digital Assistant (AIDA) [2] serves as a helper to students and is able to answer questions on course materials, generate quizzes and new learning activities and re-write the materials themselves to suit student need.

We are now beginning to bring the above together, building on [3] (also see subsection 5.2) and Tim Berners-Lee's note on 'Charlie Works'[10] to create a Lifelong Learning Coach. In particular combining GenAI, Personal Knowledge Graphs and Solid pods so that we can utilise personal student information to hyper-personalise the feedback that AIDA gives whilst preserving privacy.

References

1 Alexander Mikroyannidis, Nirwan Sharma, Audrey Ekuban, and John Domingue. Using generative ai and chatgpt for improving the production of distance learning materials. In *2024 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 188–192. IEEE, 2024.

2 Bart Rienties, John Domingue, Subby Duttaroy, Christothea Herodotou, Felipe Tessarolo, and Denise Whitelock. What distance learning students want from an ai digital assistant. *Distance Education*, pages 1–17, 2024.

[8] https://analyse.kmi.open.ac.uk/
[9] https://blockchain.open.ac.uk/
[10] https://www.w3.org/DesignIssues/Works.html

3 Piero A. Bonatti, John Domingue, Anna Lisa Gentile, Andreas Harth, Olaf Hartig, Aidan Hogan, Katja Hose, Ernesto Jimenez-Ruiz, Deborah L. McGuinness, Chang Sun, Ruben Verborgh, and Jesse Wright. Towards computer-using personal agents, 2025.

4.7 Robust explanations with Neurosymbolic AI

Michel Dumontier (Maastricht University, NL)

Language Models (LLMs) have created new opportunities to build trustworthy, transparent, and accountable AI systems. However, key challenges remain in ensuring that AI outputs can be trusted, explaining how decisions are reached, and integrating diverse sources of structured and unstructured knowledge. In our group, we explore how neurosymbolic agents that combine symbolic reasoning with machine learning techniques can be used with knowledge graphs and scientific literature to predict and explain unknown biomedical phenomena. Several projects in the group are focused on trustworthy generative AI. Our GENIUS Lab for Trustworthy Generative AI fosters academic–industrial collaboration to develop generative AI systems capable of supporting expert decision-making. Neuro-symbolic methods are used to produce interpretable reasoning capabilities, with a focus on creating open-source frameworks for conversational AI, leveraging FAIR (Findable, Accessible, Interoperable, Reusable) data and services. Complementary projects, such as REALM and CHARM, highlight collaborative efforts to harness data standards, blockchain technology, explainable AI, and advanced data management for healthcare applications. The overarching goal is to develop AI-driven solutions that not only yield accurate predictions but also provide transparent, scientifically grounded justifications. By integrating human expertise with AI-based reasoning, this research seeks to enhance the reliability, scalability, and accountability of AI systems across diverse, real-world domains.

4.8 Trustworthy Engineering of Neurosymbolic AI Systems

Fajar Ekaputra (Vienna University of Economics and Business, AT)

The rapid evolution of Neurosymbolic AI systems – particularly those that combine Knowledge Graphs (KGs) with machine learning – has opened a plethora of new possibilities for future development of AI systems. However, as these hybrid systems become more complex, they also present significant challenges. One of the most pressing concerns is the lack of standardized system representation for designing, engineering, and documenting such systems. This issue hampers the systematic characterization of these complex architectures, making them harder to analyze, compare, and trust.

To address these challenges, frameworks like boxology notation [1] have been proposed to visually simplify the representation of complex AI systems. By offering a clearer view of how different components of such systems interact, these approaches aim to improve understanding and foster greater trust. However, current solutions primarily focus on post-hoc analysis rather than geared towards supporting the entire engineering process.

In this seminar, I hope to explore ways to extend the existing AI system representations to make them more beneficial throughout the AI system development life cycle. We aim to support representation of diverse perspectives through a pattern-based engineering approach [2] – from interdisciplinary collaboration to the needs of various stakeholders and engineering processes. We also discuss how this approach could potentially enhance AI system auditability, particularly in the context of regulatory frameworks such as the EU AI Act.

References

1 Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. Modular design patterns for hybrid learning and reasoning systems. *Applied Intelligence*, 51:6528–6546, 2021.
2 Fajar J. Ekaputra. Pattern-based engineering of neurosymbolic ai systems. *Journal of Web Semantics*, 85:100855, 2025.

4.9 Using Semantic Web Technologies for Reasoning about Policies

Nicoletta Fornara (University of Lugano, CH)

Machine-readable rules and policies are fundamental to KG-based AI, as they can be used to formalize legal requirements, social norms, privacy preferences and licenses that govern the use and exchange of personal knowledge graphs between parties. For many years I have been studying systems for the formalization of norms and policies in the field of Agents and Multiagents Systems by using Semantic Web Technologies. We proposed models for representing and reasoning on obligations, by extending the ODRL language [1] and a model for representing and reasoning on norms able to generate at run-time deontic relationships [2]. Since 2021, I have been co-chair of the W3C ODRL (Open Digital Rights Language) Community Group. I coordinate the activities of the group that defines the semantics of ODRL. In this seminar I would like to investigate how languages for policy specification can be used in Knowledge Graph-based AI. I think it would be fundamental to study what types of explanations can be produced by the policies governing the statements in the various types of KG (personal KG and community KG) that are used by the ML algorithms. It is also crucial to study how to efficiently translate policies from natural language to machine readable formats and how to evaluate policies efficiently in real scenarios.

References

1 Nicoletta Fornara and Marco Colombetti. Using semantic web technologies and production rules for reasoning on obligations, permissions, and prohibitions1. *AI Commun.*, 32(4):319–334, January 2019.
2 Nicoletta Fornara, Soheil Roshankish, and Marco Colombetti. A framework for automatic monitoring of norms that regulate time constrained actions. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIV: International Workshop, COINE 2021, London, UK, May 3, 2021, Revised Selected Papers*, page 9–27, Berlin, Heidelberg, 2021. Springer-Verlag.

4.10 KG-based AI in Industrial Data Ecosystems

Sandra Geisler (RWTH Aachen, DE)

In industrial settings, such as manufacturing, data sharing is still subject to distrust. Companies rarely share their data with academia or industry partners in fear of disadvantages for their business. Highly distributed settings require new methods to utilize meaningful and reliable information from several potentially disparated and contradictory sources to be useful for higher level AI services. Examples for such settings are e.g., cross-organizational stream processing in manufacturing or informing a Digital Product Passport by distributed data analytics at suppliers. Knowledge graphs as linked and semantically rich sources provide useful information to fuel AI methods and improve the quality of their outputs. However, to utilize information from KGs, their complete life cycle and ecosystem needs to be taken into account including their generation, evolution, and integration with other KGs. Especially differing ontologies as bases for the KGs, differing goals of actors in their ecosystems, quality and provenance as well as trust between stakeholders, are still big challenges which I would like to discuss in this seminar.

4.11 AI Accountability and Data Governance from a Pragmatic Perspective.

Anna Lisa Gentile (IBM Almaden Center – San Jose, US)

With the current magnitude of AI systems, a large part of accountability depends on the traceability and availability of the data used within the model. While governance of the data starts at acquisition time, keeping complete track of the vast amount of data used for model training can be unfeasible; therefore, tools that can effectively act and correct the outputs of the models are a must-have. These tools can be thin layers of detectors that can quickly identify sensitive topics and screen unwanted output, as well as detect and quantify inadvertent leakage of proprietary data and correct the model behaviour.

4.12 Infusing Large Language Models with Knowledge: A Round-trip Ticket

José Manuel Gómez-Pérez (Expert.ai – Madrid, ES)

Data is the fossil fuel of AI. While computational power is growing with better hardware and algorithms, data is not. Large language models (LLMs) need to be pre-trained using enormous amounts of data from the Internet. However, Internet data is limited, and it is estimated that LLMs will be trained on datasets equivalent in size to the available stock of public human text data between 2026 and 2032. Therefore, pre-training as we know it will change. Additionally, when fine-tuning pre-trained LLMs for specific domains, the required training data is often unavailable, locked in corporate or regulatory silos.

To address these challenges, I propose the notion of *Knowledge as the New Fuel of AI* and a new paradigm based on infusing existing resources, such as knowledge graphs (KG), into the parametric memory of an LLM. Conversely, the knowledge contained in an LLM can also be distilled and merged into a KG, iteratively producing richer structured and parametric representations. I also introduce the concept of *Problem-Solving Prompting* (PSP) in LLMs, which builds on knowledge-based methodologies such as *Problem-Solving Methods* (PSMs) to extend prompting approaches like *Chain-of-Thought (CoT)* by breaking down complex problems into simpler subtasks.

I conclude with a call to attention on the urgent need for new benchmarks and metrics. These will be instrumental to measure both the factuality of knowledge-infused LLMs and the amount and quality of the knowledge they have been infused with.

4.13 Agreements & Accountability

Paul Groth (University of Amsterdam, NL)

Since the emergence of Large Language Models (LLMs) with the ability to perform in-context learning, we have seen a large reconsideration of knowledge engineering methods and practice. In this talk, I provide an overview of recent discussions by the community and note how LLMs shift the focus from tasks such as knowledge acquisition from text and content to other knowledge engineering tasks. In particular, the ability to come to consensus becomes paramount. How do we get assist people in coming to agreement? How do we assist people and AI systems come to agreement? I argue that knowledge graphs provide mechanism for encoding such agreement. But for that agreement to be trusted, we need to document and be able to explain how that agreement was formulated. We need agents to be accountable for the agreements that they make. Hence, I argue that we need to develop new methods for consensus formulation, mechanisms for maintaining trust, and richer approaches to explanation.

4.14 Building the Future of Enterprise AI: From Neuro-Symbolic Intelligence to Agentic Systems

Peter Haase (Metaphacts – Walldorf, DE)

The integration of symbolic and neural approaches – often referred to as neuro-symbolic AI – is gaining significant momentum, particularly in enterprise contexts where explainability, domain specificity, and trust are paramount. While neural methods such as large language models (LLMs) excel at general language understanding and pattern recognition, they are limited by a lack of transparency, factual grounding, and access to enterprise-specific knowledge. This is where symbolic technologies like knowledge graphs come into play, enabling structured, logic-based reasoning, traceability, and contextual relevance.

At the forefront of this shift is metaphacts, a company developing advanced AI capabilities through its platform metaphactory. Embracing neuro-symbolic and agentic AI, metaphacts is building metis, an intelligent agent powered by an LLM that has direct access

to a knowledge graph. Metis combines cognitive capabilities – such as natural language understanding, semantic reasoning, and planning – with execution capabilities like SPARQL query generation, semantic modeling, and retrieval-augmented generation. This enables conversational interfaces that support use cases ranging from semantic search and discovery to knowledge graph construction and ontology engineering.

These innovations are particularly relevant as the industry moves toward more mature and responsible applications of AI. While generative AI is currently experiencing a period of recalibration – having passed the peak of inflated expectations and facing challenges like hallucinations and scaling – knowledge graphs are climbing the "Slope of Enlightenment" in Gartner's Hype Cycle. They are increasingly seen as critical infrastructure for grounding AI in enterprise semantics and delivering trustworthy, explainable solutions.

Metaphacts is actively contributing to this evolution by operationalizing neuro-symbolic AI in real-world settings. For example, metis can provide transparent answers rooted in enterprise knowledge, maintain provenance, and adapt to user context, making it a powerful tool for organizations navigating regulatory requirements and complex information landscapes. Beyond retrieval, metis also assists in automating parts of the knowledge engineering process, helping reduce the cost and effort of building and maintaining knowledge graphs.

4.15 The Age of Agents: A Tale of Two Traditions

Andreas Harth (Fraunhofer IIS – Nürnberg, DE)

Introduction and Motivation. The rise of large language models has sparked new interest in agent architectures by offering powerful ways to turn natural language into actions. Such development comes as web decentralisation technologies like ActivityPub, Solid and Web of Things are maturing. These decentralised technologies preserve individual self-determination and use knowledge graphs and standardised descriptions to create environments where autonomous agents can operate.

Earlier waves of excitement about software agents in the 1990s and early 2000s, for example around shopping bots and personal digital assistants, fell short of expectations. Now that language models provide powerful means to process natural language and the infrastructure begins to support machine-interpretable representations that would allow for automated interactions with content and services, it is time to re-examine both agent architectures and their underlying technical foundations.

Classification Framework. Agent research has historically followed two main schools of thought that differ in how they view agency. The two perspectives ask fundamentally different questions: "How do we build?" versus "What does agency mean?". These perspectives can be organised as follows:

Perspective	Single-Agent Focus	Multi-Agent Focus
Technical	Mapping inputs to actions	Rationality and game theory
Philosophical/social	Nature of goals and intentions	Social interaction (simulated)

Technical Tradition. The technical view focuses on building systems using formal methods and frameworks for rational decision-making. The computational approach, developed by Nilsson/Genesereth and Russell/Norvig, sees a single agent mainly as a complex function

that turns inputs into actions. When looking at multiple agents, the tradition often focused more on theoretical aspects of rationality and game theory rather than practical challenges of distributed systems.

Philosophical and Social Tradition. The philosophical tradition, exemplified by Cohen/-Levesque, looks at basic questions about the nature of agency itself. The work explores what goals and intentions mean for artificial agents. Wooldridge/Jennings expand the view to multi-agent systems by studying the social aspects of agency, highlighting properties like independence, taking initiative, and ability to interact socially. The tradition typically used centralised simulations to study how multiple agents interact, mostly avoiding the engineering challenges of distributed systems.

Back to Basics: A Minimal Notion of Agency. One approach to advancing the field considers stripping the concept of an "agent" down to fundamental elements: an independently executing process that can be created and terminated, can communicate with other processes and can fail independently. Whether these basic units are called processes, actors or agents may matter less than examining these fundamental capabilities as potential building blocks of any agent system. More sophisticated notions of agency, whether from the technical or philosophical traditions, might then emerge from the minimal foundation.

Web Architecture as Implementation Platform. Web architecture fundamentally operates on a client/server distinction, where the client initiates interactions and the server maintains state. In addition, web architecture assumes web resources, identified via a uniform name, that represent mappings from time to representations. Multiple concurrent processes, actors or agents can be built on such an architecture by combining resources with computation. These entities operate with dual client-server capabilities, requiring both addressability and state maintenance on the server side, along with the ability to initiate interactions as clients. The basic interaction patterns follow web architecture constraints, with all communication flowing through HTTP operations between clients and servers. Such patterns align with fundamental distributed system requirements, that is, maintaining state, message-based communication and independent failure.

Current web technologies demonstrate viable implementation approaches: ActivityPub defines federation protocols for structured interactions, Solid provides personal data spaces for process state storage and Web of Things offers standardised descriptions of interaction affordances. These concrete patterns support implementing minimal distributed components within web architectural constraints.

Conclusion. The combination of language models and decentralised web technologies presents opportunities for agent systems. Examining distributed and concurrent systems principles could inform the development of foundations that address historical challenges while enabling future developments. Clear accountability mechanisms, including user-controlled policy enforcement, transparent operation logs and verifiable computation trails, will be essential for ensuring these agent systems remain answerable to their users. The path forward may lie in rigorously evaluating which elements from both agent traditions can be meaningfully implemented within the architectural constraints of the web while maintaining these accountability guarantees.

4.16 Development of New Approaches for Federated Query Processing

Olaf Hartig (Linköping University, SE)

During the Dagstuhl Seminar I wanted to work on concrete approaches related to some of the goals laid out in the proposed research agenda of the article that was the basis of the seminar [1]. In particular, related to goal DKG6 (Federated Querying), my idea was to develop an approach to consider some form of policies (e.g., access control) during the query planning process of a query federation engine. Related to goal DKG2 (Alignment with Standardized and Community Ontologies), my idea was to apply our approach to consider ontology mappings during federated query processing [2] within a concrete use case. Related to goal MRP2 (Multi-Level Policy Evaluation), I had the idea to develop a policy evaluation algorithm, and in the context of goal DI1 (Comprehensive Recording), I was considering to develop a federated query processing approach that integrates a blockchain as a federation member.

References

1 Luis-Daniel Ibáñez, John Domingue, Sabrina Kirrane, Oshani Seneviratne, Aisling Third, and Maria-Esther Vidal. Trust, Accountability, and Autonomy in Knowledge Graph-Based AI for Self-Determination. *Transactions on Graph Data and Knowledge*, 1(1):9:1–9:32, 2023.
2 Sijin Cheng, Sebastián Ferrada, and Olaf Hartig. Considering vocabulary mappings in query plans for federations of rdf data sources. In *Cooperative Information Systems: 29th International Conference, CoopIS 2023, Groningen, The Netherlands, October 30–November 3, 2023, Proceedings*, page 21–40, Berlin, Heidelberg, 2023. Springer-Verlag.

4.17 "Fundamental Science" is discovering AI

James A. Hendler (Rensselaer Polytechnic Institute – Troy, US)

Semantic Web and Knowledge Graph technology is becoming more relevant again as we see that it can help solve some of the key challenges emerging with generative AI. In this short talk, I show some challenging problems in the traditional sciences and explore why advanced AI techniques, particularly focused on data integration (and its realization as ontologies and knowledge graphs on the distributed web) are critical to making progress in these areas.

4.18 Large Language Models, Knowledge Graphs and Search Engines: A Crossroads for Answering Users' Questions

Aidan Hogan (University of Chile – Santiago de Chile, CL)

Much has been discussed about how Large Language Models, Knowledge Graphs and Search Engines can be combined in a synergistic manner. A dimension largely absent from current academic discourse is the user perspective. In particular, there remain many open questions

regarding how best to address the diverse information needs of users, incorporating varying facets and levels of difficulty. This short talk introduces a taxonomy of user information needs, which guides us to study the pros, cons and possible synergies of Large Language Models, Knowledge Graphs and Search Engines. We further present a roadmap for future research.

4.19 Building Trust in Knowledge Graphs with Provenance and Data Quality

Katja Hose (TU Wien, AT)

As the volume of data continues to grow, ensuring the reliability of AI-driven systems depends on the quality, provenance, and interoperability of the knowledge they use. Knowledge Graphs play a crucial role in structuring, integrating, and providing access to factual knowledge, which is essential for both human understanding and AI applications such as LLMs. Without mechanisms to verify and track the origins of data, errors and inconsistencies can propagate, undermining trust in AI-generated results. This is particularly relevant for Large Language Models, which often suffer from hallucinations even without relying on incomplete or incorrect information [1].

To address these challenges, validation techniques based on SHACL and ShEx help enforce structural constraints and improve the consistency of knowledge graphs [3]. Provenance tracking further enhances transparency by documenting the origins and transformations of data and how answers to queries are derived from the data [5], allowing users to assess its reliability. Efficient access to evolving knowledge and scalable mechanisms for ensuring data integrity are also crucial for maintaining trustworthy AI applications.

In addition to data quality and provenance, interoperability across different knowledge graph data models is vital. Knowledge representations vary between RDF-based graphs, property graphs, and other emerging models, making it essential to establish common foundations that support seamless integration and querying [4, 2]. Standardized schemas and transformation techniques enable interoperability, ensuring that knowledge can be effectively shared and leveraged across different systems without loss of meaning or consistency.

By strengthening provenance tracking, validation mechanisms, and interoperability strategies, we can build more reliable, accountable, and explainable AI systems. Future developments in this area will shape the way knowledge is managed and utilized, ensuring that AI-driven insights are built on a foundation of high-quality, verifiable, and interoperable information.

References

1 Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective. *Journal of Web Semantics*, 85:100844, 2025.

2 Kashif Rabbani, Matteo Lissandrini, Angela Bonifati, and Katja Hose. Transforming rdf graphs to property graphs using standardized schemas. *Proc. ACM Manag. Data*, 2(6), December 2024.

3 Kashif Rabbani, Matteo Lissandrini, and Katja Hose. Extraction of validating shapes from very large knowledge graphs. *Proc. VLDB Endow.*, 16(5):1023–1032, January 2023.

4 Shqiponja Ahmetaj, Iovka Boneva, Jan Hidders, Katja Hose, Maxime Jakubowski, José Emilio Labra-Gayo, Wim Martens, F. Mogavero, Filip Murlak, Cem Okulmus, Axel Polleres, Ognjen Savkovic, Mantas Simkus, and Dominik Tomaszuk. Common foundations for shacl, shex, and pg-schema. In *Proceedings of the ACM Web Conference 2025*, New York, NY, USA, 2025. Association for Computing Machinery.
5 Daniel Hernández, Luis Galárraga, and Katja Hose. Computing how-provenance for sparql queries via query rewriting. *Proc. VLDB Endow.*, 14(13):3389–3401, September 2021.

4.20 Remaining (Self-)Determined in a world of Agentic AI

Luis-Daniel Ibáñez (University of Southampton, GB)

In this talk I provide an overview of the challenges to self-determination that individuals face with the upcoming era of Agentic AI. AI agents could be misused to collect further data about individuals in order to nudge them towards a goal not necessarily aligned with their interests. In an extreme case, as described by Chaudhary and Penn, human motivations can be collected and sold to agents, an evolution of the economy of attention into the economy of intention. Then, I'll motivate the following research questions related to the construction of a counter-agent that help citizens navigate the economy of intention. Can an AI agent be designed to help a human retain their autonomy in a world of intent merchant agents? On what infrastructure? What is the minimum information required for such an agent to work? How much asymmetry in the intelligence of the competing agents is tolerable?

4.21 Trust and Accountability in Financial AI

Ryutaro Ichise (Institute of Science Tokyo, JP)

In this talk, I introduce two key projects related to trust and accountability.

Causality is essential for decision making in finance, yet existing knowledge graphs face challenges such as complex logical structures, inconsistencies, and limited reusability. To address these, we designed a pipeline for constructing a causal knowledge graph, FinCaKG-Onto, which integrates text, ontologies, and linked data. Our approach outperforms ChatGPT in capturing nuanced causality while avoiding generic concepts.

One application of FinCaKG is identifying dominant factors in financial causal chains. We developed a pattern mining strategy to extract dominant factors, validated through real market data. This work enhances explainability and accountability in financial decision-making.

The second project is about hallucination detection. Large Language models (LLMs) sometimes generate hallucinations – incorrect or misleading facts – which pose risks in high-stakes scenarios. Unlike LLMs, knowledge graphs offer structured reliable facts. We developed a text verification framework leveraging knowledge graphs to detect hallucinations and tested it using systematically generated hallucination data.

4.22 (Logics-aware) KG Alignment, KG Validation, KG Embeddings, Neurosymbolic AI

Ernesto Jiménez-Ruiz (City St George's, University of London, GB)

This presentation explores key aspects of Knowledge Graph (KG) technologies, focusing on: i) Knowledge Graph (KG) Alignment, ii) KG Validation, iii) KG Embeddings, and iv) Neurosymbolic AI. The talk highlights the importance of logical consistency, reasoning, and explainability in AI-driven knowledge graphs.

Knowledge Graph Alignment

Detecting Different Modeling Views. The integration of multiple models can lead to logical inconsistencies, known as unsatisfiabilities. These issues often arise due to different modeling perspectives or incorrect mappings. Algorithms for minimizing conservativity violations play a crucial role in ensuring alignment integrity.

Using Large Language Models (LLMs). Recent advancements in language models facilitate ontology subsumption inference, aiding the alignment process by identifying relationships between concepts.

Knowledge Graph Validation

Validation using Datalog Rules. Some OWL axioms are treated as integrity constraints. Missing information is captured through violation predicates, and reasoning with datalog rules ensures data consistency and completeness.

Hybrid AI: Ontology Embeddings

OWL2Vec* for Ontology Embeddings. Techniques like OWL2Vec* enable the embedding of OWL ontologies, enhancing machine learning applications by incorporating structured knowledge representations.

Learning with Knowledge Graph Embeddings. Knowledge Graph Embeddings (KGE) are utilized for classification tasks, such as predicting adverse biological effects of chemicals. These embeddings enhance explainability and facilitate reasoning over unseen entities.

Neurosymbolic AI

Learning with Prior Knowledge. Incorporating logical constraints into machine learning models helps maintain consistency by penalizing incorrect predictions. This approach bridges symbolic reasoning with data-driven AI.

Conclusion and Future Work

The presentation also provides an overview of research initiatives, including the work conducted at the AI Research Centre at City St Georges, University of London. It connects these efforts to the foundational topics of the Dagstuhl Seminar 25051, reinforcing the importance of robust validation, integration techniques, and explainability in AI-driven knowledge graphs. The research presented at the seminar emphasizes: i) The need for robust

alignment techniques to resolve inconsistencies in multi-source KGs. ii) The role of reasoning mechanisms in ensuring logical validity. iii) The integration of embeddings and symbolic AI for improved decision-making. Further exploration in neurosymbolic AI and adaptive KG alignment mechanisms will be crucial for advancing trust and accountability in AI systems.

4.23 Threats to Trust in Organizations' Operationalized Knowledge

Timotheus Kampik (SAP Berlin, DE & Umeå University, SE)

In my presentation, I discussed (classes of) threats affecting the trustworthiness of operationalized knowledge that is used to define how socio-technical systems run in and across organizations. I argued that threats relating to malicious behavior (attacks) and lack of compliance (e.g., privacy issues) are relatively well-understood. In contrast, there are more subtle classes of threats that are prevalent, severe, and poorly understood. Specifically, bullshit knowledge emerges when (human) agents are compelled to formalize knowledge they do not fully understand or care about, and that formalized knowledge is typically not generalizable, meaning that its trustworthiness depends on context. How these threats can be addressed in a systematic manner remains to be seen.

4.24 Compliance Technologies for Trust

George Konstantinidis (University of Southampton, GB)

In this talk I present recent advances in data compliance technologies. I discuss how the use of computational and machine processable policy languages can enable data usage control, encoding a range of rules from legislation to regulation and environmental or other preferences. I discuss how compliance algorithms implemented on policy engines can verify compliance or detect conflicts. I present our tools for managing, tracking and updating user consent and preferences on data and AI operations and pipelines. I present an approach for automated negotiation and execution of policies, contracts and agreements. Lastly, I present a framework and roadmap for trust and reputation management algorithms and systems.

4.25 Neurosymbolic GeoAI

Manolis Koubarakis (University of Athens, GR)

It has been shown by Tony Cohn and coauthors (COSIT 2024) that current large language models do not perform well on spatial reasoning problems. For example, they cannot answer questions such as "You are walking south along the east shore of a lake and then turn around to head back in the direction you came from, in which direction is the lake? Is it to your

left or to your right?" (Correct answer: Left; LeChat by Mistral does not answer correctly; ChatGPT answers correctly). My current research concentrates on evaluating the most recent large reasoning models on spatial reasoning tasks such as the above, and combining LLMs and spatial reasoners for solving such tasks effectively (hence, the neurosymbolic term in the title). In this way, we will be able to develop chatbots that will do better in geographic tasks and be more useful than current ones in geospatial applications (e.g., way-finding).

4.26 Knowledge Graph-Aware AI in the Evolving AI Landscape

Deborah L. McGuinness (Rensselaer Polytechnic Institute – Troy, US)

We are living in an age of rapidly advancing technology. History may view this period as one in which generative artificial intelligence is seen as reshaping the landscape and narrative of many technology-based fields of research and application. Times of disruptions often present both opportunities and challenges. I briefly introduce some areas for discussion about how and where knowledge graphs (both personal KGs and other KGs) may be positioned in emerging hybrid architectures that may provide value propositions that might impact adoption. I also provide some dimensions (such as explainability, interoperability, etc , through which we may view and evaluate the potential of knowledge graphs in today's landscape.

4.27 Machine processable policies for socio-technical systems

Julian Padget (University of Bath, GB)

Policies are one important source of trust for entities participating in socio-technical systems because they offer guarantees on accountability as well as expectations of behaviour and the achievability of goals. My past work on norm representation and reasoning has taken a formal approach using action languages, which was then combined with ODRL to provide an operational semantics for fragments of GDPR represented in ODRL. Other contributory contextualising factors for trust, although out of scope here, but still relevant are various standards and guidelines for process and for technologies, such as IEEE 7001-2021, IEEE 7003-2024, ISO 42001 and the UN Guide on Privacy-Enhancing Technologies. My goals for this seminar are to explore more effective ways to build and maintain machine processable policies, facilitated by large language models, but alongside formal approaches, to support the engineering of socio-technical systems.

4.28 Trusting Query Results

Philipp D. Rohde (TIB – Hannover, DE)

Federated query processing answers a query retrieving data from multiple sources as if they were a single source. This requires (semantic) source descriptions as well as query decomposition and planning with respect to the capabilities of the different sources. The data within a KG can be validated against (integrity) constraints using shape-based languages like SHACL or ShEx. When it comes to process-based data, e.g., cancer patients following the treatment guideline, a new shape-based validation language, PALADIN, is proposed. But when it comes to trust in query results, different perspectives need to be considered. A computer scientist might have a different view on what is trust than the average user. The quality of the data is only one dimension that contributes to trust. Other dimensions like access control and provenance are also discussed.

4.29 Trust-Based Decision Support Systems

Daniel Schwabe (Rio de Janeiro, BR)

We investigate how hybrid systems integrating Knowledge Graphs (KGs) and generative language models assist decision-making in various domains. Our goal is to explore how these systems can best support decision processes. The decision-making process should be understood as a reasoning mechanism that aligns with the intended goals (purpose) of a human agent executing a particular action, taking into account their personal preferences, characteristics, and values. To reach a decision, the agent assesses Its decision policies applied to trusted information. This includes both circumstantial information about the situation at hand and contextual information on relevant factors. To obtain trusted information, the agent accesses various sources, including potentially crowdsourced Knowledge Graphs (KGs) and Large Language Models (LLMs). The agent then applies its trust policies to the information retrieved from these sources to extract reliable information. To trust a particular piece of information, the user constructs (possibly recursively) a trust chain of claims, evidence, and supporting proofs to make a final trust decision. To decide if a particular claim is to be used as a fact for the specific intended action, there are three possible alternatives.

1. The agent already accepts that claim as a fact because it already knows it to be the case;
2. The claim is to be accepted as a fact because of social norms. For example, it was made by an agent with public faith, such as a notary public;
3. If all resources have been exhausted, e.g, time, computational resources, lack of additional information, etc., the agent makes an arbitrary decision which is locally referred to as a "leap of faith". In other words, accepting a claim as a fact without any kind of evidence.

Our research investigates the various architectures, representations, and functionalities of hybrid systems (also called neurosymbolic systems) that can support this decision-making. Specifically, we are looking on how to include explicit representations of context in knowledge graphs, and how to support justification dialogues, using both knowledge, graphs, and LLMs in supporting the decision process, for specific domains.

4.30 Bridging Resilient Accountable Intelligent Networked Systems (BRAINS)

Oshani Seneviratne (Rensselaer Polytechnic Institute – Troy, US)

This short talk introduces the research conducted at the BRAINS Lab at RPI, which focuses on enhancing trust and accountability in decentralized AI systems. Our work spans three high-level areas: decentralized privacy-preserving data infrastructures, smart contract innovations, and foundation model innovations with decentralized technologies. At its core, this research explores how decentralized knowledge graph ecosystems can empower individuals while ensuring safety, transparency, autonomy, and alignment with human values. In the context of this Dagstuhl Seminar, I am particularly interested in advancing several foundational topics. For machine-readable policies and norms, I aim to explore how knowledge graphs can accurately represent policies and how these policies can remain adaptable and enforceable in decentralized systems. For decentralized infrastructures, I focus on how best to ingest knowledge into vertical and hybrid federated learning systems and the design of knowledge-infused architectures for LLMs. For decentralized knowledge graph management, I seek to address the challenges of sustainably managing decentralized ecosystems by keeping knowledge up-to-date, incentivizing contributions and verification, and handling contradictions or diverse viewpoints. Finally, for explainable neuro-symbolic AI, I am investigating how to ensure AI safety and protect personal data when integrating LLMs with personal knowledge graphs while providing clear and effective explanations.

4.31 Trust, Accountability, and Autonomy in Generative Health AI

Chang Sun (Maastricht University, NL)

This talk addresses the dimensions of trust, accountability, and autonomy in the development and deployment of generative AI systems for health data. It introduces a novel methodology for generating synthetic patient data for rare or previously unseen diseases using ontology-enhanced generative adversarial networks (Onto-CGAN). By integrating biomedical ontologies into the training process, the approach improves the quality and relevance of synthetic data, enabling machine learning models to generalize more effectively in scenarios where real data is scarce. While synthetic data does not fully match the performance of real-world data, it significantly outperforms models trained on limited or no data.

In addition to unimodal synthetic data generation, the talk explores the application of multimodal language models to radiological visual-linguistic tasks, highlighting the need for interpretability and task-specific evaluation in clinical settings. The presentation also introduces the ciTIzen-centric DAta pLatform (TIDAL), a privacy-preserving infrastructure designed to support dynamic and fine-grained digital consent management. Built on Solid (SOcial LInked Data) principles and employing the Data Privacy Vocabulary, TIDAL enables secure, decentralized storage and governance of personal health data, with consent-aware federated learning capabilities. All data and consent artifacts are represented in RDF, allowing for semantic interoperability and standards-based integration.

4.32 You can't pin down trust, but you can still do something

Aisling Third (The Open University – Milton Keynes, GB)

Concepts like trust get defined in various fields, without necessarily representing the same phenomenon. This carries the risk of serious problems where, e.g., technical systems with different definitions interact. This is by no means unique to trust, of course, but we can observe that its nature as a foundational concept of social interaction makes it easier for unconscious assumptions to come into play. It is clear that we need to handle these concepts of trust in a flexible way. This sort of problem is what the Semantic Web can be useful for: interoperability by making concepts explicit. But it is equally unlikely that formal languages can be used to capture these concepts either. We argue therefore that it would be more fruitful to model instead the factors which go into making trust decisions, e.g., user ethical and social values, any requirements of secrecy, etc., including how these relate to trust, so that operationalising trust decisions can still be handled by relevant actors with the required information to do so.

4.33 The value of trust that surrounds data

Ruben Verborgh (Ghent University, BE)

For the longest time, we assumed that Linked Data – public or private – was about technologies that facilitate the transfer of data. Maybe we had it all wrong. When we download RDF from DBpedia, we essentially get back a list of triples. Like any series of bytes, these map losslessly to a list of natural numbers and back. With the natural numbers being a known – albeit infinite – set, DBpedia cannot possibly send us any new data points. Hence, it must be sending us something else. The story becomes very different when we realize that the value of what DBpedia sends us, is not in the numbers themselves, but in the trust assessment that DBpedia is implicitly making when sending them. Namely: this is a list of triples to which DBpedia attaches some truth value. And while there similarly exist an infinite number of such lists, we attach value to this particular one, because DBpedia endorses it. Unfortunately, we as a community are not very explicit at all about the semantics of that trust, which makes it hard to capture and discuss value. Let's talk about trust.

4.34 Hybrid AI Systems with Knowledge Graphs: Enabling Trust, Accountability, and Autonomy

Maria-Esther Vidal (TIB – Hannover, DE)

Artificial Intelligence (AI) is transforming science and medicine by enabling powerful predictive and decision-support systems. However, ensuring trust, accountability, and autonomy in AI remains a challenge, particularly when models operate as black boxes. Hybrid AI

systems, combining symbolic reasoning with machine learning, offer a promising approach to overcoming these challenges. This presentation discusses the integration of Knowledge Graphs (KGs) into neuro-symbolic AI systems, emphasizing their role in enhancing interpretability and ensuring robust decision-making. Knowledge Graph (KG) ecosystems provide structured, semantic representations of knowledge, supporting data integration, constraint validation, and symbolic reasoning. A KG ecosystem is defined by various components, including data sources, ontologies, mappings, and constraints, facilitating the construction of AI systems that are both explainable and trustworthy. The life cycle of KG-based AI systems involves services, actors, roles, constraints, and requirements that ensure sustainable and transparent AI-driven decision-making. Hybrid AI leverages both symbolic and neural components. Symbolic methods, such as constraint validation and rule-based reasoning, ensure valid and explainable link prediction and counterfactual inference. Meanwhile, neural components, including numerical learning, KG embedding models, and large language models (LLMs), enhance learning from unstructured data while preserving logical consistency. The synergy between these components enables the development of AI systems capable of self-determination, improving autonomy in critical applications. The discussion will explore principled vs. integrated neuro-symbolic systems, highlighting the need for AI architectures that ensure reliability without sacrificing efficiency. This principled approach fosters trust in AI-driven applications, particularly in domains requiring high levels of interpretability, such as medicine and scientific research.

4.35 Towards neuro-symbolic agents that represent legal entities on the Web

Jesse Wright (Open Data Institute – London, GB)

The notion of agentic AI is seeing resurgent popularity in the age of LLMs-based AI. We pose a research agenda towards building hybrid agents which use LLMs to provide a human interface for agents and to support the negotiation capability of agents – whilst query and reasoning is used to provide operational safeguards to data "belief" and "sharing".

To support the maintenance of proof and provenance over derived data that agents receive we propose directions including zero knowledge proofs or e.g. SPARQL query correctness, to enable agents to have models for establishing whether the provenance they receive is sufficient to take that data to be true. We propose personalised trust modelling so agents can learn what sources their users are willing to take as authoritative for particular tasks. We also propose personalised privacy preference modelling to enable agents to automate data sharing.

5 Breakout Groups

5.1 Machine Readable Norms and Policies

Piero A. Bonatti (University of Naples, IT)
Irene Celino (CEFRIEL – Milan, IT)
Andrea Cimmino (Polytechnic University of Madrid, ES)
Nicoletta Fornara (University of Lugano, CH)
Andreas Harth (Fraunhofer IIS – Nürnberg, DE)
Luis-Daniel Ibáñez (University of Southampton, GB)
Timotheus Kampik (SAP Berlin, DE & Umeå University, SE)
George Konstantinidis (University of Southampton, GB)
Julian Padget (University of Bath, GB)
Oshani Seneviratne (Rensselaer Polytechnic Institute – Troy, US)

Abstract. As AI systems increasingly mediate complex interactions in socio-technical ecosystems, the need for formal, machine-readable representations of norms and policies becomes critical. This report introduces the concept of Computational Policy Languages and formalizes core policy reasoning tasks: generation, activation, evaluation, and enforcement. We define a Policy Engine as a computational artifact capable of supporting these tasks against a dynamic, knowledge-graph-based representation of the world. To ground this framework, we explore four operational scenarios – intending, attempting, monitoring, and auditing – that structure the temporal and procedural dimensions of policy application. A set of diverse use cases, including business process compliance, financial contract management, industrial safety, organizational governance, and energy data sharing, illustrates the breadth of challenges and requirements such systems must address. This report consolidates a research agenda for formal, interoperable, and context-aware policy systems, identifying open problems at the intersection of logic, semantics, system design, and regulatory alignment.

5.1.1 Introduction

Complex ecosystems for Knowledge-Graph based AI include multiple interactions between their participants. In several scenarios, a participant's action may hurt the self-determination of another (human) actor. For example, an AI agent actor deciding to deny a benefit, potentially in an unlawful manner, without providing the right to recourse. Another example is a human actor sharing some data that was considered private or sensitive by another actor.

To remediate, it is essential to equip these ecosystems with the means to define and enforce norms, policies, and rules so they can express 1. Global norms that all actors are expected to follow, or rules that their actions must not break. This could serve, for example, to encode principles that protect the self-determination of all actors. 2. Individual rules, constraints, or preferences of actors establishing boundaries with respect to actions from other actors.

A desirable characteristic of these policy languages is being machine-readable and machine-processable. The reason is two-fold: first, it facilitates the checking of when a policy has been violated by an action, or if an individual preference is incompatible with a global norm; second, when AI-agents are actors in the ecosystem, allow them to read the policies.

- Trust: Clear definitions of rules that every actor can evaluate and compare following a deterministic algorithm improve trust in the whole ecosystem.
- Accountability: In combination with the appropriate log and trace systems, determine which actor has violated a rule and what the consequences or repair actions are.
- Autonomy: On the one hand, rules can be designed to protect the autonomy of certain actors, while on the other, their existence allows AI agents to become more autonomous, in the sense they can perform more actions with confidence a rule is not violated.

In this short paper, we provide a review of the state of the art, our proposed approach based on the definition of computational policy languages, a list of use cases from the practice of the participants' group discussion, including desiderata and challenges for computational policy languages to support them.

5.1.2 State of the Art

Policy Languages were initially designed to tackle the problem of *Access Control* and its generalisation to *Role-based access control* (RBAC): An organisation defines a number of roles, to be fulfilled, an actor or agent playing a role needs access to data and resources whose access is governed by policies. An RBAC model and implementation must guarantee that users can access the required data and resources in accordance with organisational policies [4].

The notion of Access Control was further generalised to *Usage Control* [13]. Access Control can be considered as a problem of Authorisation or Permission to perform an action on a target object, and usage control adds the concepts of obligations and conditions. Obligations are requirements that have to be fulfilled for usage allowance. Conditions are environmental or system requirements that are independent of individual subjects and objects. Usage control also emphasizes the continuity of enforcement. Policies are enforced not only before access but also during and after the agent is acting upon the target object. If attributes change during access and the policy is no longer satisfied, usage control systems may revoke the granted access and terminate the usage.

Ibáñez et al. [8] classifies policy languages as general and specific. General languages cater to a diverse range of functional requirements (e.g., access control, query answering, service discovery, negotiation), whereas specific languages focus on a single functional requirement. A number of general languages were developed, but none of them achieved mainstream adoption. On the specific languages front, the Open Digital Rights Language (ODRL), which is a W3C recommendation, has gained a lot of traction in recent years thanks to its use to express intellectual property rights management. Additionally, the ODRL model and vocabularies have been extended to model contracts, personal data processing consent, and data protection regulatory requirements.

Akaichi and Kirrane [2] defines a usage control framework as a complete solution that allows for the specification of usage control policies, the enforcement of said policies, and the realization of policies and enforcement mechanisms via a usage control system. They provide a taxonomy of requirements for a usage control framework. The three top categories are (i) Specification, representing requirements relating to policy expressiveness, defined semantics, as well as flexibility and extensibility of the policy language; (ii) Enforcement, or mechanisms used to enforce and manage usage policies throughout the usage process, which consists of three phases: before usage, ongoing usage, and after usage; and (iii) System, that refers to non-functional requirements such as usability and performance.

A problem of additional interest is the application of Usage Control in a decentralised system, or a complex scenario with multi-agent systems.

Kampik et al. [9] discusses the relevance of norms, policies, and preferences for governing complex sociotechnical multiagent systems on the Web. The key challenge they identify is the integration of normative concepts with WoT abstractions and systematic evaluation of the practical usefulness of the integration results. They propose a conceptual framework that serves to define the role played by various norms, policies, and preferences when it comes to complex sociotechnical systems on the Web and demonstrate it via a simple but realistic scenario.

[11] consider the general distributed case; they propose a generic and formal model that allows for the explicit distinction of different systems, their individual behaviors, as well as their interplay, enabling reasoning about the distributed system they form. The first model and implementation that transparently and generically tracks dataflows and policies across systems.

5.1.3 Proposed Approach: Computational Policy Languages

We propose the following definition of Computational Policy Languages:

▶ **Definition 1** (Computational Policy Language)**.** A Computational Policy Language is a formal language used to define, reason with, and enforce policies, including permissions, prohibitions, and obligations, to control and govern the usage of resources and behavior of actors in socio-technical systems.

▶ **Definition 2** (State of the World)**.** A data structure that holds knowledge about the state of the socio-technical systems on which Computational Policies apply. In the spirit of the seminar, we assume the State of the World is encoded in a Knowledge Graph. Depending on context of application, the state of the world may be a snapshot of the system a ta given time, or contain information about states across a time continuum.

Based on these definitions, we identify the following scientific problems associated *Policy Computational Languages*

▶ **Definition 3** (Policy Generation)**.** In general, this problem refers to the translation of Natural Language to a Computational Language. We recognise the following variations of the problem:

- Policy creation/authoring/specification: given a description of a policy in natural language, generate a policy in a computational policy language. It is also possible to consider the problem of creating a User Interface to generate policies.
- Policy legitimation: given a machine-readable or machine-processable format, generate a natural language version expressed with domain-specific terms, such that it is admissible in a given legal framework.
- Policy explanation: given a machine-readable, machine-processable, or natural language policy, generate a natural language description of what the policy is about.

▶ **Definition 4** (Policy Activation)**.** Given a state of the world and a list of policies, select the subset of policies that are relevant to be evaluated against the state of the world. For example, if a policy states that the Actor must be inside the library on Tuesdays between 09:00 and 18:00 and the state of the world is: Today is Wednesday, 10:00, and the actor is in the Library, then the policy is not 'active' in this state of the world.

▶ **Definition 5** (Policy Evaluation). Given a Computational Language Policy and a state of the world, decide if the state of the socio-technical system violates the policy or not. For example, if a policy states that Actor must be inside the library on Tuesdays between 09:00 and 18:00, then for the following states of the world:

- Is Tuesday 10:00 and the actor is in the Library: The state does not violate the policy.
- Is Tuesday 10:00 and the actor is in the Kitchen. The state violates the policy.

▶ **Definition 6** (Policy Enforcement). Given a state of the world and a set of violated policies, compute a set of updates to the state of the world such that the updated state of the world does not violate the input set of policies.

Finally, for practical reasons, we consider the definition of an artefact that encapsulates algorithms for each of the problems, that we dub *Policy Engine*.

▶ **Definition 7** (Policy Engine). A system that integrates algorithms that solve each of the scientific problems of a Computational Policy Language.

5.1.3.1 Operational Scenarios

The operationalisation of the solutions to the problems described in section 5.1.3 depends on the contextual scenario in which they are invoked. We identify four *operational scenarios* for the application of policies.

- *Intending* refers to when an agent intends to execute an action on the state of the world, before proceeding, the agent asks a policy engine to activate and evaluate policies to understand if any violation would occur.
- *Attempting* refers to when an agent attempts to execute an action on the state of the world, before letting the action affect the state of the world, a policy engine activates and evaluates policies to decide if any violation would occur.
- *Monitoring* refers to a process or meta-system that monitors the actions multiple agents execute upon the state of the world, continuously activating and evaluating policies upon those actions.
- *Auditing* refers to the post-facto analysis of a temporal trace of actions upon the state of the world with the purpose of determining if any action on the trace violated any of the policies active at the moment it was executed.

5.1.4 Use Cases and Challenges

In this section, we describe five use cases that would benefit from Computational Policy Languages, highlighting their desiderata and what challenges they pose to the computational problems defined in section 5.1.3.

5.1.4.1 Business Process Conformance Checking

Process conformance evaluates whether an executed business process conforms to certain policies or norms. For example, in an internal audit, an organization may want to check to what extent employees engage in so-called *maverick buying*, *i.e.*, the issuing of a purchase requisition before the approval of the purchase. Maverick buying cannot generally be prohibited. For example, employees may need to be able to purchase work equipment even when their line management is not available and hence cannot issue approvals before the fact. However, if maverick buying is rampant, this indicates that employees are exploiting the system, undermining cost control.

Operational scenario/life-cycle steps. Process conformance checking is a well-established method in business process management, describing the assessment of how well traces of real-world process execution conform to expected behavior [5]. Such expected behavior can be derived, for example, from external regulations, internal best practices, or knowledge about system fundamentals (for data quality checks). In the life cycle of business process management, which in its simplest form consists of process design, execution, and analysis, conformance checking is typically assigned to the analysis step. Accordingly (and in accordance with the maverick buying example), the most common operational scenario is *auditing*, followed by *monitoring*, when applied to cases of a process that have not yet terminated.

Computational problems. The primary computational problem in conformance checking is to determine whether a trace of a process execution is conformant or not. This corresponds to *policy evaluation* in an auditing scenario. However, other computational problems are relevant as well. For example, comparing (sets of) conformance rules helps assess to what extent different organizational conformance requirements agree or, on a more fundamental level, are logically consistent with each other.

Desiderata/challenges. A suitable computational policy language needs to fulfill the following high-level requirements.

- **Logical time as first-class abstraction.** In business process management, logical time is generally considered a crucial first-class abstraction and is commonly formalized using Petri nets [16] or abstractions utilizing finite-trace linear temporal logic [6]. Accordingly, a policy language that can be used or designed for conformance checking must feature a notion of logical time.
- **Operationalization on (Big) Symbolic Data.** In practice, conformance checks are typically executed on large amounts of data that is often stored and queried using special-purpose database systems [10, 17]. Accordingly, policies for process conformance must have an operational semantics that ultimately allows for execution in the context of the aforementioned systems, or by mainstream (e.g., SQL-based) query engines.
- **Models of agents and their roles.** In business processes, (human and software) agents interact, executing activities in order to achieve an organizational objective. Accordingly, the notion of an *agent* is important. In business process management, agents are traditionally called *resources* that have specific *roles* when executing activities [18]. More recent work features the notion of an *agent*, more specifically in a meta-model for process traces that can be utilized for agent-oriented process conformance checking [14].
- **Meta-level Meaning/Labels of Policy Rules.** When a process trace violates a compliance rule, the nature of the rule impacts the implication that non-conformance has. For example, rule violation may imply (on the meta-level) that the process trace does not comply with a specific regulation, is likely to negatively impact organizational performance, or even that the underlying data most be incorrect (e.g., if according to the trace, a message is received before it is sent) [1]. Accordingly, a policy language for process conformance must support the labeling of policy rules according to their meta-level meaning.
- **Support for Deontic Notions.** Some academic work on process conformance checking applies deontic logic [7]. Still, it is not clear how exactly mainstream process conformance checking relates to classical deontic notions of permission, prohibition, and obligation. However, conformance checking typically distinguishes between *imperative* (*How does the process have to behave?*) and *declarative* (*How is the process permitted to behave?*) approaches, thus reflecting deontic ideas.

5.1.4.2 Over the Counter Financial Derivatives Contracts

Over-the-counter (OTC) financial derivatives contracts are customized agreements between counterparties, and their terms must be explicitly defined in a machine-readable format. The contract terms – such as payment schedules, interest rate adjustments, margin requirements, and credit events – are encoded as rule-based policies. OTC contracts can have customized tenors (durations) ranging from days to decades. This flexibility requires a computable policy that governs time-sensitive aspects, including contract lifecycle management, expiration triggers, and time-dependent risk adjustments. OTC derivatives contracts are inherently flexible, allowing for bilateral negotiations and secondary market trading. They can be transferred to new counterparties through assignment or novation while retaining the original policy conditions or incorporating additional terms.

The OTC contracts consist of multiple phases from trade initiation to termination, with several operational steps (Depicted in Figure 1).

1. The terms of the contract (e.g., notional amount, tenor, strike price, underlying asset) are agreed upon bilaterally for the *trade initiation*.
2. This is followed by *trade confirmation* where the contract details are matched, and signed by the counterparties.
3. *Collateral exchange* then takes place to mitigate the counterparty risk.
4. During the period the OTC contract takes place, several *trade lifecycle events*, such as interest payments, happen.
5. When the contract reaches the expiration date *trade termination* happens.

In the case of a transfer of the OTC contract between `Party A` and `Party B`, by `Party A` to another counterparty `Party C`, we may encounter the following set of activities, given the nature of the transfer.

1. *Assignment (Partial Transfer):* The original counterparty `Party A` assigns its rights and obligations to a new counterparty `Party C`. Party A is still legally responsible unless explicitly released by the original agreement.
2. *Novation (Full Transfer)*: The entire contract is legally replaced with a new agreement. The original party (`Party A`) is fully discharged, and the new party (`Party C`) takes full responsibility.

When transferring an OTC derivative, the core contract structure remains unchanged unless explicitly renegotiated. In other words, obligations, risk exposure, and collateralization remain as originally defined. In some cases, counter-parties may wish to modify the contract when trading it to a new party. Additional conditions can include credit enhancement clauses (e.g., requiring a third-party guarantor for counter-parties with a lower credit rating), trigger-based clauses (e.g., automatic termination if market conditions exceed predefined risk thresholds), or regulatory compliance adjustments (e.g., reporting structure adjustments due to jurisdiction changes).

Desiderata/challenges. A suitable computational policy language needs to fulfill the following high-level requirements.

- Possibility to define obligations, risk, exposure, and definition of collaterals.
- Ability to model policy validity and expiration as a function of time.
- The concept of a contract as a first-class citizen. An interesting question is whether Smart Contract languages can be classified as computational policies or if they should be regarded as an extension with additional problems.
- Support for a single policy involving multiple actors with different obligations.

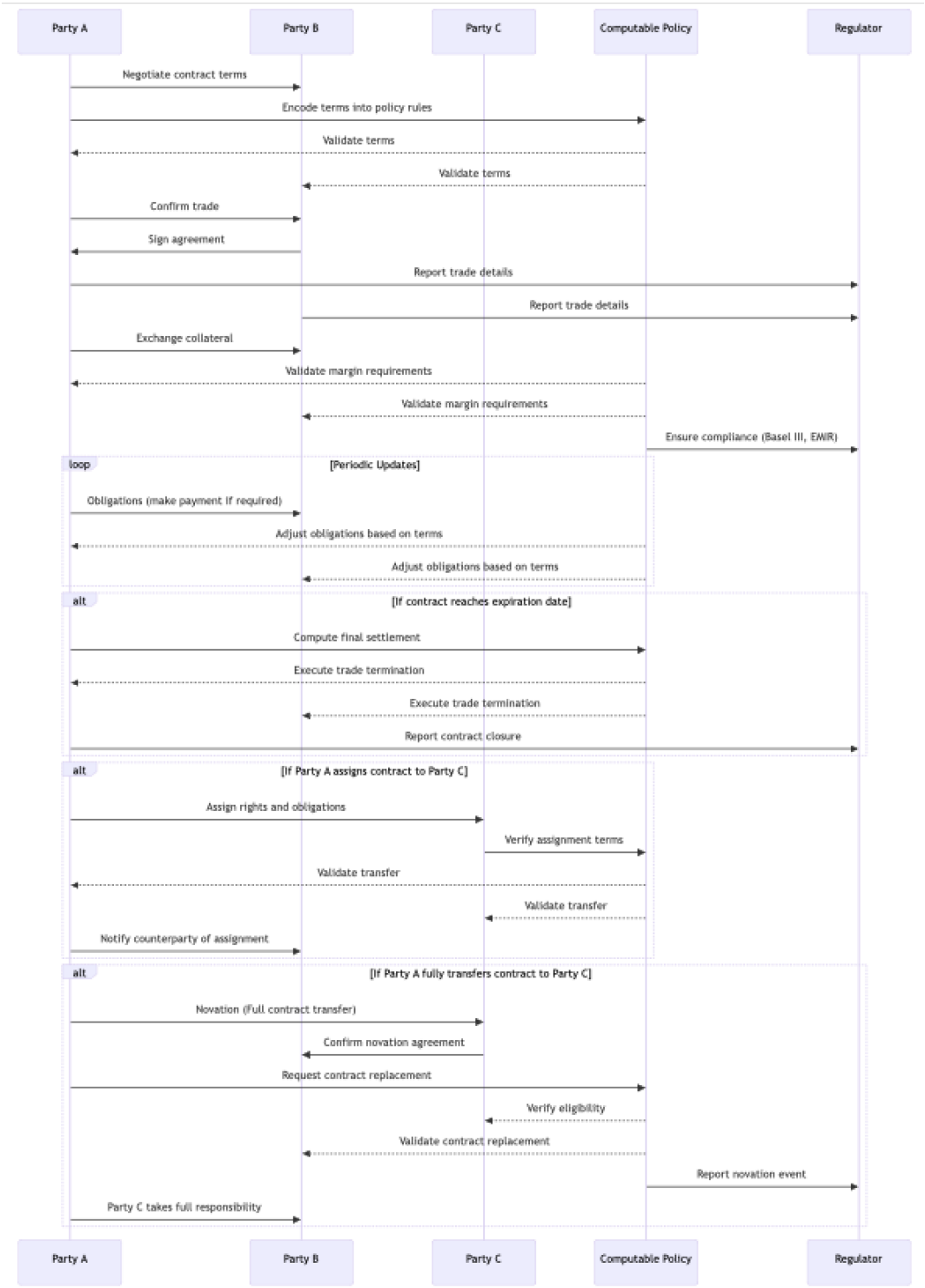


Figure 1 Sequence Diagram of Policy Decisions in OTC Financial Derivatives Contracts.

- Embed compliance rules from regulations such as Basel III, Dodd-Frank, or European Market Infrastructure Regulation (EMIR) to ensure legal adherence in an auditing operational scenario.
- The problem of enforcement usually involves fund transfer, imposing additional security challenges.

5.1.4.3 Industrial Safety Policies

In a company, especially but not limited to manufacturing companies, employees are required to follow *safety procedures* when operating on a production line, e.g., to make a maintenance intervention. Procedures exist to guide expert users to follow the expected process and ensure compliance. In parallel to operational procedures, *policies* and *guidelines* are usually provided to regulate the expected behaviours.

In this scenario, the concept of policy is interpreted as the commitment of one party to behave to match an expectation of another party: expectations may concern future states of affairs or future behaviour of the parties involved; the policy is used to track fulfillment and violation, rather than specific actions or steps to be executed. Therefore, by industrial safety policies we refer to the strategic aspect of safety (opposed to the operational aspect of safety procedures); in other words, within the same company the same policy (strategic safety) can be valid across different procedures (operational safety): for example, different lines within the same factory may need different operational steps due to the differences in the machines, but they must all be compliant to the same expected (strategic) safety behaviours.

Policy Management Life-cycle Steps. All steps, but especially intention, monitoring and auditing. For *intention*, *employees* are interested in checking if a specific behaviour is expected, mandatory, or prohibited, e.g., wearing protective equipment in specific contexts. For *execution/monitoring*, *operators* want to make sure that the production line is constantly in a safe state, while *controllers* are interested in checking that all employees behave safely to ensure the smooth operations of the line. For *auditing*, violations of safety behaviours must be collected to intervene and prevent future misbehaviours.

Computational problems. Mainly activation/evaluation and enforcement. When checking the policy to get informed about the allowed/prohibited behaviours, *activation/evaluation* takes place, to select all policies that apply to the specific state of the world (e.g., the specific activities going on in the factory) and to verify that the expected behaviours match the expectations and, in case they don't, indicate that there is a violation. When such a violation is identified, *enforcement* is needed to understand the consequences (e.g., corrective action, formal/official warning notice, etc.). Across those problems, some opportunities and challenges may emerge to apply *inference* and provide *explanations*: the factory "state of the world" may be multi-faceted, making it complex to identify the applicability of a policy (e.g., employee role, production line, type of intervention on the machinery, temporal and spatial context, etc.); at auditing time, repeated cases of safety policy violation by an employee may lead to specific sanctions or to the need to identify and apply corrective actions, e.g., training/retraining courses.

Desiderata/challenges. A policy language to fulfill the above scenario should take into account the following requirements:

- Ability to correctly formulate and check expected behaviours, especially in the case of concepts like safety that may be overlooked and considered "common sense" knowledge.

- Possibility to precisely define roles and their accountability/responsibility w.r.t. the strategic policy: in the case of safety, there are clear social and legal implications.
- Alignment between company-specific policies and legislation in force: policy specification needs to make sure that "local" permissions/prohibitions are not in contrast with laws.
- Interplay between (strategic) policies and (operational) procedures: if strict compliance with specific processes is crucial, policies should also include the procedural knowledge (see previous scenario); otherwise, policies must ensure that the expected behaviors are in line with operational indications.

5.1.4.4 Business Travel Expenses Reimbursement

Description. In any company where employees travel to perform part of their work, the need arises to manage and reimburse the travel expenses incurred by each person. While there are policies specific to each organization, there are usually documents explaining the rules for reimbursement, which may include a definition of eligible expenses, spending limits, and temporal deadlines for the reimbursement process management. An *employee* is interested in understanding if an expense can be reimbursed, sometimes even before making a purchase, while *responsible managers* and *financial departments* are interested in checking the correctness of reimbursement claims and acting upon them. Whenever a problem arises during the reimbursement process, the involved actors are also interested in understanding what happened and why, to perform remedial actions (policy enforcement).

Policy Management Life-cycle Steps. All steps, but especially intention, monitoring and auditing. For *intention*, employees are interested in checking if an expense will be eligible for reimbursement before making the payment. For *attempting/execution*, employees submit their claims for expense reimbursement and wait for their processing. For *monitoring*, the finance department may wish to supervise the various requests for reimbursement and have a list of actions to be carried out within specific time constraints. Similarly, employees would like to know the list of reimbursements they have to submit and the corresponding time scales. For *auditing*, all actors want an explanation on the expenses that were not approved and understand how to act about it, and managers specifically may also be keen to check the behaviours of their employees to identify the need or obligation of sanctions, as well as recognition for constant fulfillment.

Computational problems. Activation, evaluation and enforcement. (i) Checking the policies to get informed on their state of *activation*, to select all policies that regulate expense reimbursement and potentially those that apply to the specific state of the world (e.g., the specific type of expense a given employee incurred in). (ii) When checking for eligibility of reimbursement, the issue of *evaluation* arises to verify if the expenses match the expectations and, in case they do not, indicate that there is a violation. (iii) When such a violation is identified, *enforcement* is needed to understand what the steps to be followed are (e.g., reject an expense that overcame the spending limit, communicate to the user, etc.). In addressing these issues, some opportunities and challenges may arise for the application of *inference* techniques and for providing *explanations* for the reasoning performed: the user "state of the world" may be articulated and precise to understand the applicability of a policy and correctly "instantiate" it (e.g., employee level, type of expense, temporal and spatial context, etc.) especially in presence of a high number of exceptions and corner cases; with specific reference to trust, repeated cases of policy violation by a reimbursement requester may lead to specific sanctions or to the decrease of the trust other actors have w.r.t. them.

Desiderata/challenges. A policy language to fulfill the above scenario should take into account the following requirements:

- Bridge between the natural language version of the policies and potentially automated systems to process the computational version.
- Consider the dimension of trust, as trust can change between the parties when policies are evaluated and enforced.
- Consider the distinction between the deontic part of the policy (*regulative rules*, e.g., permissions/prohibitions/obligations/rights of expenses) and the descriptive/definition part of the policy (*constitutive rules*, e.g., what means that an expense is business-related, what means that an expense is excessive, etc.). There is, from the representative prospect, a continuum between those two parts (as policies may mix "schema and instances"), and Knowledge Graphs seem to be a fitting solution to cover their representation.
- Management of exceptions, especially those usually managed "by hand" on a case-by-case basis.
- Support the policy dynamics over time, as this kind of regulation may change unpredictably due to business, organizational, or legislative reasons.

5.1.4.5 Energy System Data Sharing

Description. In 2023, the UK Government commissioned a report to examine the case for a data-sharing infrastructure (DSI) for the UK energy system [3]. A follow-up assessment [12] makes the case for the urgent development of an MVP to explore the needs of influencing and impacted stakeholders in a diversified energy system for which a whole-system approach becomes essential for its effective management, in contrast to the current siloed, hierarchical structure. We consider operational issues such as data cleaning, etc., out of scope for this discussion, although data quality – and hence the policies governing it – pervade such a system and clearly embody risk to system function. However, here we focus on the many machine-processable policies that can facilitate the function and the evolution of the data-sharing infrastructure, in contrast to fundamentally fragile solutions that rely on top-down, regimented control with mandated representations.

Policy Management Life cycle Steps. All steps are critical for this scenario. *Intending* and *attempting* matter for a participant in the DSI so that both they and the DSI can be assured ahead of time that an action is compliant with the policies active at the time and that an action will not affect – as can best be determined – system integrity. *Monitoring* is a logical follow-on that observes (sequences of) actions to assess the continuing performance of the system as a whole and how individual actions are contributing (or not) to the achievement and maintenance of system goals. System operating actors may then, in line with additional policies, step in to alter the system trajectory and keep it within the desired behavioural envelope. Lastly *auditing* serves to process the record of participants' actions to contextualise individual actions against the bigger picture at the time, such as in the case of retrospective analysis of incidents and accidents, but also for the system operator to uncover behavioural patterns at scale that may indicate the need for policy revisions or additional policies to maintain system performance, and to carry out functions for the daily balancing market.

Computational problems. These are largely the same as in the other use cases, but in contrast to the definitions in Section 5.1.3 the state of the world is likely to be decentralised rather than a single data structure and hence partially observable for participants, while the system operator may have a complete but not necessarily up-to-date representation

of the state of the world. Policy generation will primarily be under the control of the system operator, although some features of the activation of a policy may subject to the requirements of the parties governed by a particular instantiation of a policy, for example, in the case of what features are shared and the privacy enhancing technology [15] to be used for sharing. Enforcement will be the responsibility of the system operator, using the monitoring and auditing mechanisms to obtain evidence of what happened when. Some aspects of enforcement may be enacted by system software actors, but others may transfer to human actors representing organisations participating in the DSI for resolution by human governance mechanisms.

Desiderata/challenges. Identified technical challenges [12] include a lack of common standards adoption by organizations in the sector, a lack of scalable infrastructure, and a prevalence of inflexible legacy systems. Associated cultural challenges [12] include perceived value of private data that inhibits sharing, concerns over the data quality of others, and potential embarrassment over an organization's own data quality.

At first sight, a regimented solution appears to offer simplified governance, but in reality, it pushes the governance burden on to

(a) data producers

(b) data consumers

(c) system maintenance.

The last is a hidden cost and a threat to evolution: over time, possibly even quite rapidly, the *one-size-governs-all* approach will meet an incompatible use case. The possible resolutions are to reject the use case, force the use case into the existing framework, or change the framework, but inertia will generally work against the last option. Thus, the challenge here is to embed sufficient flexibility in the policy framework such that is captures a space of acceptable policy solutions, which in turn implies the existence of over-arching meta-policies – these could be formal and represented in a policy language, or in natural language interpreted by humans, or a mix of both – that constrain actual policies, while themselves also being changeable to account for shifts in system requirements and participant values over time.

5.1.5 Conclusion

This report outlines a structured foundation for advancing the study of machine-readable norms and policies within knowledge graph-based AI systems. By formalizing the concept of Computational Policy Languages and introducing precise definitions for policy-related tasks – generation, activation, evaluation, and enforcement – we aim to enable rigorous reasoning about normative systems in complex, multi-agent socio-technical environments.

The proposed framework situates policy reasoning within a dynamic socio-technical system or *state of the world*, modeled as a knowledge graph. It also introduces the Policy Engine as a computational artifact to operationalize core reasoning tasks. These tasks are contextualized through four operational scenarios: *intending*, where agents assess policy compliance before acting; *attempting*, where actions are evaluated just prior to state changes; *monitoring*, which continuously evaluates ongoing behavior; and *auditing*, which retroactively analyzes actions against historical policy states.

The use cases examined in this report – spanning process conformance, financial contract execution, safety compliance, organizational governance, and energy data sharing – highlight the interdisciplinary nature of the problem space. They raise critical research challenges, such as handling temporal and deontic logic, reconciling declarative and procedural representations, modeling multi-party obligations, ensuring semantic alignment between natural language and formal policies, and integrating domain-specific ontologies and regulatory constraints.

Addressing these challenges requires a combination of formal methods, natural language understanding, semantic web technologies, and socio-legal modeling. Future research must also investigate the trade-offs between expressiveness and tractability, the integration of explanation and accountability mechanisms, and the adaptation of policy reasoning to decentralized and federated systems.

By consolidating key problems and illustrating their practical implications, this report invites further investigation into principled, interoperable, and context-aware policy languages – paving the way for a new generation of AI systems respectful of the self-determination of their actors.

References

1 Greta Adamo, Stefano Borgo, Chiara Di Francescomarino, Chiara Ghidini, Nicola Guarino, and Emilio M. Sanfilippo. Business process activity relationships: Is there anything beyond arrows? In Mathias Weske, Marco Montali, Ingo Weber, and Jan vom Brocke, editors, *Business Process Management Forum – BPM Forum 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings*, volume 329 of *Lecture Notes in Business Information Processing*, pages 53–70. Springer, 2018.

2 Ines Akaichi and Sabrina Kirrane. A comprehensive review of usage control frameworks. *Computer Science Review*, 56, 2025.

3 Arup, Energy Systems Catapult, and University of Bath. Digital spine feasibility study: exploring a data sharing infrastructure for the energy system, 8 2024.

4 Elisa Bertino, Piero Andrea Bonatti, and Elena Ferrari. Trbac: a temporal role-based access control model. In *Proceedings of the Fifth ACM Workshop on Role-Based Access Control*, RBAC '00, page 21–30, New York, NY, USA, 2000. Association for Computing Machinery.

5 Josep Carmona, Boudewijn F. van Dongen, Andreas Solti, and Matthias Weidlich. *Conformance Checking – Relating Processes and Models*. Springer, 2018.

6 Claudio Di Ciccio and Marco Montali. Declarative process specifications: Reasoning, discovery, monitoring. In Wil M. P. van der Aalst and Josep Carmona, editors, *Process Mining Handbook*, volume 448 of *Lecture Notes in Business Information Processing*, pages 108–152. Springer, 2022.

7 Laura Giordano, Alberto Martelli, and Daniele Theseider Dupré. Temporal deontic action logic for the verification of compliance to norms in asp. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, page 53–62, New York, NY, USA, 2013. Association for Computing Machinery.

8 Luis-Daniel Ibáñez, John Domingue, Sabrina Kirrane, Oshani Seneviratne, Aisling Third, and Maria-Esther Vidal. Trust, accountability, and autonomy in knowledge graph-based ai for self-determination. *Transactions on Graph Data and Knowledge*, 1(1), 2024.

9 Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian 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 Technology*, 22(4):1–31, 2022.

10 Timotheus Kampik and Cem Okulmus. Expressive power and complexity results for signal, an industry-scale process query language. In Andrea Marrella, Manuel Resinas, Mieke Jans, and Michael Rosemann, editors, *Business Process Management Forum – BPM 2024 Forum, Krakow, Poland, September 1-6, 2024, Proceedings*, volume 526 of *Lecture Notes in Business Information Processing*, pages 3–19. Springer, 2024.

11 Florian Kelbert and Alexander Pretschner. Data usage control for distributed systems. *ACM Transactions on Privacy and Security*, 21(3):1–32, 2018.

12 Furong Li, Nigel Turvey, Lewis Dale, John Scott, Julian Padget, Isaac Flower, Jennifer R. Fitzpatrick, Nico Ostler, Rob Oldaker, and Simon Yeo. Do we need a data sharing infrastructure for the energy sector? *IET Smart Grid*, n/a(n/a).

13 Jaehong Park and Ravi Sandhu. The ucon-abc usage control model. *ACM Transactions on Information and System Security*, 7(1):128–174, 2004.

14 Qingtan Shen, Artem Polyvyanyy, Nir Lipovetzky, and Timotheus Kampik. Agent system event data: Concepts, dimensions, applications. In Wolfgang Maass, Hyoil Han, Hasan Yasar, and Nicholas J. Multari, editors, *Conceptual Modeling – 43rd International Conference, ER 2024, Pittsburgh, PA, USA, October 28-31, 2024, Proceedings*, volume 15238 of *Lecture Notes in Computer Science*, pages 56–72. Springer, 2024.

15 United Nations (Statistics Division). *The United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*, 2.

16 Wil M. P. van der Aalst, Marlon Dumas, Chun Ouyang, Anne Rozinat, and Eric Verbeek. Conformance checking of service behavior. *ACM Trans. Internet Techn.*, 8(3):13:1–13:30, 2008.

17 Thomas Vogelgesang, Jessica Ambrosy, David Becher, Robert Seilbeck, Jerome Geyer-Klingeberg, and Martin Klenk. Celonis PQL: A query language for process mining. In Artem Polyvyanyy, editor, *Process Querying Methods*, pages 377–408. Springer, 2022.

18 Petia Wohed, Wil M. P. van der Aalst, Marlon Dumas, Arthur H. M. ter Hofstede, and Nick Russell. On the suitability of BPMN for business process modelling. In Schahram Dustdar, José Luiz Fiadeiro, and Amit P. Sheth, editors, *Business Process Management, 4th International Conference, BPM 2006, Vienna, Austria, September 5-7, 2006, Proceedings*, volume 4102 of *Lecture Notes in Computer Science*, pages 161–176. Springer, 2006.

5.2 Towards Computer-Using Personal Agents

Piero A. Bonatti (University of Naples, IT)
John Domingue (The Open University – Milton Keynes, GB)
Anna Lisa Gentile (IBM Almaden Center – San Jose, US)
Andreas Harth (Fraunhofer IIS – Nürnberg, DE)
Olaf Hartig (Linköping University, SE)
Aidan Hogan (University of Chile – Santiago de Chile, CL)
Katja Hose (TU Wien, AT)
Ernesto Jiménez-Ruiz (City St George's, University of London, GB)
Deborah L. McGuinness (Rensselaer Polytechnic Institute – Troy, US)
Chang Sun (Maastricht University, NL)
Ruben Verborgh (Ghent University, BE)
Jesse Wright (Open Data Institute – London, GB)

Abstract. Computer-Using Agents (CUA) enable users to automate increasingly-complex tasks using graphical interfaces such as browsers. As many potential tasks require personal data, we propose Computer-Using Personal Agents (CUPAs) that have access to an external repository of the user's personal data. Compared with CUAs, CUPAs offer users better control of their personal data, the potential to automate more tasks involving personal data, better interoperability with external sources of data, and better capabilities to coordinate with other CUPAs in order to solve collaborative tasks involving the personal data of multiple users.

5.2.1 Introduction

Advances in Generative AI, and particularly Large Language Models (LLMs), have led to the recent release of various *Computer-Using Agents* (*CUAs*) that automatically operate a user's computer on their behalf. These agents use multimodal capabilities to interact with graphical interfaces via simulated mouse and keyboard inputs. Prominent commercial examples of CUAs include OpenAI's Operator, Google's Jarvis, and new functionalities in Anthropic's Claude.

Potential use cases for CUAs involve personal and often sensitive data, such as credit card details for purchases, passport numbers for flight booking, addresses for deliveries, and allergy information for dinner reservations. While modern browsers sometimes store personal data to autocomplete web forms, CUAs could additionally take context into account (e.g., selecting between a home or work address, depending on the purchase) and go beyond simple autocompletion.

Passing personal data to CUAs raises valid concerns about how such data might be (mis)used. Currently, OpenAI's Operator invokes a *takeover mode* for tasks involving sensitive data (e.g., log-in or payment details): the user is required to fill the details in manually [25]. Such measures target users' concerns about how their personal information will be used by CUAs. OpenAI themselves state that Operator is *"still learning, evolving and may make mistakes"* [25]. There are thus many open questions relating to the use of personal user data by CUAs.

Conversely, there are many potential benefits to users if CUAs are empowered with personal data. CUAs could autofill forms with personal data for users in a context-aware and potentially generative manner, automating a tedious task. CUAs could potentially enrich personal data with public data to better solve tasks. The CUAs of multiple users could negotiate to achieve a mutually beneficial result based on their users' personal context and preferences.

Towards providing users more oversight over their personal data while enabling higher levels of automation for complex tasks, we propose **Computer-Using Personal Agents** (**CUPAs**): *a Computer-Using Agent (CUA) that has controlled access to a structured repository of private information relating to a user.* This concept is illustrated in Figure 2. Specifically, we propose to instantiate the repository as a *Personal Knowledge Graph* (*PKG*) representing the user's personal data, which would facilitate the specification by users on how the CUA can access and use these data. This PKG can collect more personal data over time, with policies also evolving to reflect the user's fluctuating trust in the system [2]. Looking further forward, one can then imagine a scenario where CUPAs interact with websites and services via the underlying Web APIs instead of through a vision model, where CUPAs can assist in recommendations and negotiations based also on interactions with similar users and/or users' CUPAs.

We provide a road-map towards realising this vision of CUPAs, discussing what is achievable now with current technology, and what gaps must be addressed via further research and development.

5.2.2 User Scenario

Sam is expecting Jane over for dinner at 8pm, and is thinking about preparing Thai food. Sam is pre-diabetic, while Jane has a shellfish allergy. Sam requests that his CUPA to generates some suggestions of Thai recipes for the occasion. Consulting Sam's schedule, the CUPA recommends to filter recipes requiring more than an hour to prepare based on

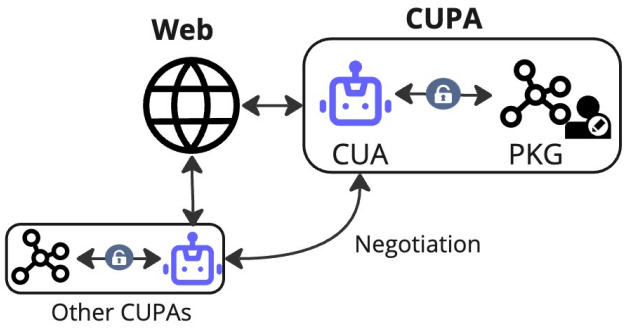


Figure 2 Computer-Using Personal Agent.

when he finishes work and his commute time. Sam agrees, and the CUPA starts to retrieve and present shellfish-free recipes of Thai food that are quick to prepare. Upon consulting external sources of nutritional information and recipes on the Web, the CUPA flags some recipes as being above the postprandial glucose threshold recommended by Sam's doctor (<180 mg/dL), or as having high glycaemic indices (>70).

Sam asks his CUPA to find out what recipe Jane might like. As Sam and Jane are friends, Sam's CUPA can send the candidate recipes to Jane's CUPA to see what she might like. Jane's CUPA suggests to avoid some recipes that include coriander (listed in some recipes as cilantro), which Jane hates. Sam's agent enforces his glucose thresholds and flags ingredients with high glycaemic indices, using external food and recipe knowledge graphs (e.g., the FoodKG [16]) to find the alternative ingredients. Of the remaining options, Sam's agent suggests a tofu green curry recipe that catches Sam's eye. Since the recipe is flagged for having a high glycaemic index (78), the agent asks Sam if he might consider replacing jasmine rice with cauliflower rice as a healthier option. Sam refuses the substitution as it is a special occasion.

Sam requests his CUPA to order the ingredients from a local supermarket. Since green bell peppers are out of stock, the CUPA suggests to replace them with yellow bell peppers. Sam agrees, and the CUPA prepares the order for delivery to Sam's home address, soliciting Sam's confirmation. Later that night, Sam and Jane enjoy their dinner of Thai green curry. After Jane leaves, Sam suffers some slight heartburn. He requests his CUPA to order antacids and additionally registers the fact that green curry dishes may cause Sam heartburn for future reference.

5.2.3 State of the Art

Personal data play an increasingly important role in modern life [24, 6]. Early works [18] characterise such data based on the *concept of six senses*: owned by me, about me, directed to me, sent by me, already experienced by me, and useful to me. More restrictive definitions only include data created by the individual [3], or that the individual cares for/about [11, 12].

Much literature has been dedicated to Personal Information Management systems (PIMs), which deal with the acquisition, organisation, maintenance, retrieval and sharing of personal data [19]. Notable PIM technologies include blockchain systems [36, 14], systems capturing user behaviour on multiple user devices [22, 26], and end-user prototypes [9, 24, 20]. Personal Knowledge Graphs (PKGs) [7, 8, 27] further apply a graph abstraction to personal data, opening up possibilities for declarative access policies, deductive inference, and integration with external Knowledge Graphs.

Towards taking fuller advantage of such data, AI-powered agents show much promise, particularly those that can automate tasks currently performed by the user. Robotic Process Automation (RPA) [29, 13] automates interactions with human interfaces. However, such approaches are hard-coded, brittle to changes in the interface, and incapable of generalising to unseen interfaces. Conversely, AI-based agents are capable of learning and generalising. LLM-based agents have been proposed to operate in diverse environments using recursion, feedback, and careful prompt engineering [33]. Such LLM-based agents are capable of solving computer tasks – despite the limited reasoning capabilities in LLMs [21] – paving the way for CUAs such as Operator [25].

Regarding works unifying LLM-based agents with PKGs, AGENTIGraph [34] heads in this direction, but rather focuses on question answering. Closer to the idea of CUPAs is Charlie: a brief proposal by Berners-Lee [4] on combining LLM-based agents with PKGs instantiated by Solid pods using Semantic Web standards. This proposal, and the user scenario presented previously, echo the (yet unrealised) vision laid out by Berners-Lee et al. [5] for the Semantic Web 24 years ago. Wright [31] presents a "discuss then transact" model of LLM-interaction in support of this vision for LLM-based personal agents that represent legal entities.

5.2.4 Added Value

Societal and legal debates on personal data emphasise *protection* from the harm that they could inflict, and understandably so. Yet people voluntarily exchange personal data with others in their every-day lives in the pursuit of mutual benefit. People can decide to leverage more personal data, or different kinds of personal data, to achieve a desired outcome. For instance, patients might prefer to share fitness-tracker data with their doctor if this improves their treatment, or consumers might want to divulge allergies and dietary needs to streamline online shopping and avoid nasty surprises.

A dangerous assumption is that companies are more capable of distilling value from people's personal data than the people the data describe. A company certainly has advantages over individuals in this respect, such as the ability to aggregate over a great many users. But personal data about a particular individual in isolation has much greater potential to empower that individual than a company they interact with, especially when the individual is coached by an agent such as a CUPA. CUPAs representing different parties could even negotiate a better outcome for *all* parties involved.

Considering the added value of CUPAs, and more generally of providing AI-based agents access to personal data, we highlight:

Multi-dimensional negotiation. CUPAs can help users to strike sweet-spots between multiple dimensions, such as the cost and duration of multi-hop flights, the deliciousness and healthiness of meal options, etc.

Increased granularity. Humans struggle to negotiate on a fine-grained level, and may thus prefer broad policies that reduce cognitive load (e.g., to always accept all cookies) [32]. CUPAs can help to reach fine-grained agreements that improve outcomes and honour party preferences.

Improved risk/reward assessment. CUPAs can help users simulate and analyse a variety of hypothetical data exchange scenarios, and warn users of a particular risk, for example that the supermarket – if informed of a condition of a severe allergy – could sell this information to third parties, leading to an increase in life assurance premiums.

Auditing and follow-up. CUPAs could automatically perform audits to assess whether the data were treated as agreed during the negotiation process, evaluate the benefit to the user, and improve for future interactions.

Such added value is, of course, dependent on the value outweighing the potential harms caused. This can be addressed via AI alignment, which ensures that artificial intelligence systems act in accordance with human intentions, values, and societal norms. It involves *outer alignment*, where an AI's objectives accurately reflect human goals, and *inner alignment*, ensuring learned behaviours remain aligned in novel scenarios. Machine-readable policies on how personal data from the PKG can or should be used by the AI-based agent can also help to avoid harm. Representing personal data as PKGs allows standards such as the Open Digital Rights Language (ODRL) [17] and policy engines implementing formal semantics [15] to specify and automate the processing of policies about how personal data are used, in what contexts, and under what conditions.

5.2.5 CUPA Capabilities

Computer-using personal agents must be able to *interact with diverse websites and APIs*. This allows them to book flights and hotels, search for job openings, and even schedule appointments. Moreover, they must possess the ability to *interact with other such agents*, such as coordinating travel arrangements with a travel agent or collaborating with a financial agent to manage expenses.

In addition to being able to *generate and adapt content* (e.g., personalised summaries and creative text), a computer-using personal agent must be able to *combine private data* from the user's personal knowledge graph (PKG) with external information. For example, when searching for a new apartment, the agent should combine the user's preferred neighbourhood from their PKG with data from real estate websites and local amenities databases to find the most suitable options. When utilising the knowledge stored within the PKG, the agent must also be able to *adapt the knowledge from the PKG for the current task*. For instance, when filling out a job application form, the agent should selectively use information from the user's CV and work history stored in the PKG, tailoring the presentation to the specific requirements of each application. This adaptability is crucial for ensuring that agent actions are relevant and effective in the given context.

CUPAs must continuously *collect and enrich user information* to effectively assist them. This involves gathering data from various sources, including interactions with websites and APIs, user inputs, and external sources. By continuously *learning about user preferences*, these agents can personalise their assistance, such as recommending travel options that align with the user's preferences or suggesting recipes that cater to specific dietary restrictions or tastes. However, it is also crucial for such agents to *avoid learning one-off or irrelevant patterns*, for example, to assume that Sam will always suffer heartburn after eating Thai food and should thus avoid it.

Computer-using personal agents must exhibit a high degree of autonomy. They should ideally *act maximally autonomously*, including the ability to *proactively anticipate and address user needs*. For example, an agent could proactively remind users of upcoming appointments or suggest relevant articles based on their recent reading history. However, this autonomy must always be balanced with the ability to *be guided and controlled by the user*, allowing users to provide instructions, adjust preferences, and maintain control over agentic actions.

While acting largely autonomously, it is crucial that a computer-using personal agent *acts in alignment with the user*, ensuring that tasks are completed as desired. This is essential in scenarios like recipe searches where the agent must accurately reflect dietary restrictions and preferences. Moreover, such an agent should always act in the user's interests, even when *dealing with potentially conflicting goals*. For example, an agent helping a user plan a trip should consider factors like budget, travel time, and personal preferences, even if these factors

may lead to a slightly more expensive or less convenient option. The agent should avoid *acting in an unethical or illegal manner* even if it potentially maximises a users immediate interests, e.g., via tax evasion.

To maintain user trust and ensure responsible behaviour, it is also crucial that agents *do not overstep bounds*, respecting user privacy and only acting within explicitly granted permissions. Finally, the repeated offering of *clear explanations of all actions* will aid in the fostering of trust and allow users to understand and verify agent behaviour.

5.2.6 Technical Challenges

The aforementioned desired capabilities for CUPAs, based on our vision of a trusted, accountable and largely autonomous agent acting with personal data for user benefit, raises a number of technical challenges.

Accountability and Liability In the case of undesired, illegal, or unethical acts involving CUPAs, it is important to determine who – or what – is responsible, who should be held accountable, and where the liability lies.

Explainability, Traceability, and Provenance Provenance techniques are required to trace and explain how personal and external data led to specific answers or actions being derived or carried out by the CUPA. These provenance techniques would need to support diverse data models, machine learning processes, user inputs and policies.

Data Interoperability Data interoperability is a key challenge towards implementing CUPAs. Being able to draw on and integrate more sources of data will improve the CUPAs performance. This is particularly challenging for new sources discovered on the fly.

Inter-Agent Communication, Negotiation and Coordination Agents must communicate effectively in the context of multi-agent systems to achieve shared goals, requiring both a shared conceptual understanding and a means of encoding and decoding messages [30]. The same challenge applies to networks of CUPAs who coordinate to solve a particular set of goals for users.

Security, Privacy, and Policies The sensitive nature of data processed by a CUPA calls for security, privacy, and usage control mechanisms, and the ability of the CUPA to understand and correctly apply the access/usage/action control policies of the user. In some countries, this would even be a legal requirement (e.g., under GDPR in the E.U.).

Trust, Delegation, and Action Control Achieving agent autonomy requires trust modelling, delegation mechanisms, and structured action control policies [28]. Trust models must be adaptable to different contexts, from rigid policies applicable in government agencies to more flexible, reputation-based approaches for personal agents [10].

User-in-the-Loop CUPAs will require input, guidance, permission and confirmation from the user. But to increase automation, the CUPA must avoid unnecessary interactions with the user. This creates the challenge of *when* to call upon the user, and how.

Self-Improvement The CUPA should leverage its experience with the user in order to improve the services it provides over time, leading to greater automation, and actions/results that better benefit the user. This raises questions about how such a history can be captured, represented, stored and leveraged.

Self-Determination and Alignment

5.2.7 Roadmap

We envisage that moving from the current state of the art to fully addressing the above technical challenges will occur in three stages. These levels represent varying degrees of trust, accountability and autonomy.

CUAs enhanced with personal data In the first instance, we foresee extensions of CUAs – in the style of OpenAI's Operator [25] in a commercial setting and Agent-E [1] in a research setting – such that they use a PKG in order to access knowledge personal to the user. This would safely enable higher levels of automation, obviating the need to pass control back to the user in scenarios of the user's choosing that involve personal data.

Web-aware CUPAs CUAs currently rely on existing browser implementations to render an HTML page and then make use of vision models to interact with the page. An agent could rather observe HTTP requests made to a particular website, as well as the HTML forms present on a page, to invoke requests and actions directly via HTTP.

Networks of CUPAs We envision networks of CUPAs interacting in order to complete tasks involving multiple users. This may involve structured service descriptions [23], or a mix of natural language and structured communication per a "discuss then transact" model [31] whereby agents use natural language to first negotiate about a transaction they wish to perform, and then confirm this transaction using structured data.

5.2.8 Conclusion

Computer-Using Agents (CUAs) have the potential to transform how users interact with their computers, their browsers and amongst themselves. Not having access to personal data limits such interactions. Giving CUAs unfettered access to the personal (and most sensitive) data of a user seems unwise, as does providing CUAs no access to personal data. We thus argue for CUPAs as a configurable middle-ground, where a Personal Knowledge Graph (PKG) is used to represent, store and control access to the personal data of the user. As a starting point, the data that a user fills into web forms can be captured in the PKG, and enriched by an AI-based agent. These data can then be used, if the user so wishes, by CUAs to automate further tasks. In a next step, CUPAs can learn to interact with websites via HTTP APIs rather than though visual interfaces. Finally, we envisage further into the future a network of CUPAs collaborating to address users' tasks.

References

1 Tamer Abuelsaad, Deepak Akkil, Prasenjit Dey, Ashish Jagmohan, Aditya Vempaty, and Ravi Kokku. Agent-e: From autonomous web navigation to foundational design principles in agentic systems. *arXiv preprint arXiv:2407.13032*, 2024.

2 Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1–30, 2024.

3 Ofer Bergman and Steve Whittaker. *The science of managing our digital stuff*. MIT Press, 2016.

4 Tim Berners-Lee. Charlie Works. Design Issues, `https://www.w3.org/DesignIssues/Works.html`, 2025.

5 Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

6 Kean Birch, D. T. Cochrane, and Callum Ward. Data as asset? the measurement, governance, and valuation of digital personal data by big tech. *Big Data and Society*, 8, 2021.

7 Prantika Chakraborty, Sudakshina Dutta, and Debarshi Kumar Sanyal. Personal research knowledge graphs. In *WWW 2022 – Companion Proceedings of the Web Conference 2022*, pages 763–768. Association for Computing Machinery, Inc, 4 2022.

8 Prantika Chakraborty and Debarshi Kumar Sanyal. A comprehensive survey of personal knowledge graphs, 11 2023.

9 Amir Chaudhry, Jon Crowcroft, Heidi Howard, Anil Madhavapeddy, Richard Mortier, Hamed Haddadi, and Derek McAuley. Personal data: Thinking inside the box. *Aarhus Series on Human Centered Computing*, 1:4, 2015.

10 Ray Chen, Fenye Bao, and Jia Guo. Trust-based service management for social internet of things systems. *IEEE transactions on dependable and secure computing*, 13(6):684–696, 2015.

11 Amber L. Cushing. PIM as a caring: using ethics of care to explore personal information management as a caring process. *Journal of the Association for Information Science and Technology*, 74(11):1282–1292, 2023.

12 Amber L. Cushing and Páraic Kerrigan. Personal information management burden: A framework for describing nonwork personal information management in the context of inequality. *Journal of the Association for Information Science and Technology*, 73:1543–1558, 11 2022.

13 Diogo António da Silva Costa, Henrique São Mamede, and Miguel Mira da Silva. Robotic Process Automation (RPA) adoption: a systematic literature review, 6 2022.

14 Benedict Faber, Georg Michelet, Niklas Weidmann, Raghava Rao Mukkamala, and Ravi Vatrapu. Bpdims:a blockchain-based personal data and identity management system. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2019-Janua:6855–6864, 2019.

15 Nicoletta Fornara, Víctor Rodríguez-Doncel, Beatriz Esteves, Simon Steyskal, and Benedict Whittam Smith. ODRL Formal Semantics, May 2024.

16 Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness, and Mohammed J. Zaki. FoodKG: A semantics-driven knowledge graph for food recommendation. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web – ISWC 2019*, pages 146–162, Cham, 2019. Springer International Publishing.

17 Renato Iannella and Serena Villata. ODRL Information Model 2.2, Feb 2023.

18 William Jones. *The future of personal information management, part 1: Our information, always and forever*. Morgan & Claypool Publishers, 2012.

19 William P Jones and Jaime Teevan. *Personal information management*, volume 14. University of Washington Press Seattle, WA, 2007.

20 Varvara Kalokyri, Alexander Borgida, and Am lie Marian. YourDigitalSelf: a personal digital trace integration tool. *International Conference on Information and Knowledge Management, Proceedings*, pages 1963–1966, 2018.

21 Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language Models can Solve Computer Tasks. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 – 16, 2023*, 2023.

22 Jiangxu Lin and Meng Wang. PKG : A Personal Knowledge Graph for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11â • fi15, 2022, Madrid, Spain*, volume 1, pages 3334–3338. Association for Computing Machinery, 2022.

23 David Martin, Massimo Paolucci, Sheila McIlraith, Mark Burstein, Drew McDermott, Deborah McGuinness, Bijan Parsia, Terry Payne, Marta Sabou, Monika Solanki, Naveen Srinivasan, and Katia Sycara. Bringing semantics to web services: The OWL-S approach. In Jorge Cardoso and Amit Sheth, editors, *Semantic Web Services and Web Process Composition*, pages 26–42, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

24 Richard Mortier, Jianxin Zhao, Jon Crowcroft, Liang Wang, Qi Li, Hamed Haddadi, Yousef Amar, Andy Crabtree, James Colley, Tom Lodge, Tosh Brown, Derek McAuley, and Chris Greenhalgh. Personal data management with the databox: What's inside the box? In *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking*, CAN '16, page 49–54, New York, NY, USA, 2016. Association for Computing Machinery.

25 OpenAI Team. Introducing Operator. OpenAI Blog `https://openai.com/index/introducing-operator/`, published 2025-01-23, accessed 2025-01-29, 2025.

26 Markus Schröder, Christian Jilek, and Andreas Dengel. A Human-in-the-Loop Approach for Personal Knowledge Graph Construction from File Names. In *Knowledge Graph Construction*, volume 3141. CEUR Workshop Proceedings, 2022.

27 Martin G Skjæveland, Krisztian Balog, Nolwenn Bernard, Weronika Łajewska, and Trond Linjordet. An ecosystem for personal knowledge graphs: A survey and research roadmap. *AI Open*, 5:55–69, 2024.

28 Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. Authenticated delegation and authorized ai agents. *arXiv preprint arXiv:2501.09674*, 2025.

29 Wil M.P. van der Aalst, Martin Bichler, and Armin Heinzl. Robotic process automation. *Business and Information Systems Engineering*, 60:269–272, 8 2018.

30 Michael Wooldridge. *An introduction to multiagent systems*. Wiley, 2009.

31 Jesse Wright. Here's Charlie! Realising the semantic web vision of agents in the age of LLMs. *CoRR*, abs/2409.04465, 2024.

32 Jesse Wright, Beatriz Esteves, and Rui Zhao. Me want cookie! Towards automated and transparent data governance on the Web, 2024.

33 John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 – 16, 2023*, 2023.

34 Xinjie Zhao, Moritz Blum, Rui Yang, Boming Yang, Luis Márquez Carpintero, Mónica Pina-Navarro, Tony Wang, Xin Li, Huitao Li, Yanran Fu, Rongrong Wang, Juntao Zhang, and Irene Li. AGENTiGraph: an interactive knowledge graph platform for LLM-based chatbots utilizing private data, 2024.

35 Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment. *Philosophical Studies*, November 2024.

36 Guy Zyskind, Oz Nathan, and Alex Sandy Pentland. Decentralizing privacy: Using blockchain to protect personal data. *Proceedings – 2015 IEEE Security and Privacy Workshops, SPW 2015*, pages 180–184, 2015.

5.3 Evaluation of AI Systems

Paul Groth (University of Amsterdam, NL)
Michel Dumontier (Maastricht University, NL)
Michael Cochez (VU Amsterdam, NL))
Fajar Ekaputra (Vienna University of Economics and Business, AT)
Monica Palmirani (University of Bologna, IT)

Abstract. The evaluation of AI systems is central to fostering trust and ensuring accountability. A systematic assessment of such systems offers insights into a model's strengths and limitations regarding its performance on specific tasks. Furthermore, rigorous evaluation reveals how well the model generalizes across different scenarios, handles uncertainty, and adheres to ethical standards. Traditional AI systems evaluation focuses on the benchmarking approach using task-based evaluation metrics concerning ground truth [2].

However, such evaluations are fraught with challenges due to the lack of benchmark datasets, difficulties in creating gold standards, and the complexity of assessing new problems and domains. In this working group, we outline these challenges and propose innovative strategies for evaluating AI systems.

5.3.1 Discussed Problems

Evaluating complex AI systems presents unique challenges. These challenges have been highlighted in recent meta-reviews of evaluation failures across machine learning systems, such as those discussed by Liao et al. [1], emphasizing issues like implementation variations, overfitting, and metrics misalignment. Some of the key challenges are in the following:

1. Lack of Benchmark Datasets: Creating gold standards and benchmark datasets for new problems and domains is a difficult task. Even harder is to create a benchmark that is such that when a systems performs well on it, then the system can be applied on a context broader than the one the benchmark was derived from.
2. Dynamic Contexts: For example, in the legal domain, changing legislation and environments can render benchmark datasets obsolete.
3. Complexity of Domains: Each domain, such as public policy or clinical trials, has unique contexts that complicate standard evaluation approaches.

In addition to these challenges, several common pitfalls can lead to misleading results when evaluating AI systems. Some issues are data-related: *Incomplete training and test splits* can cause unexpected distribution shifts between training and deployment, leading to poor generalization. *Missing or incorrect ground truth data* can result in inconsistent or unbalanced datasets, often due to insufficient annotators or domain expertise. Another critical problem is *data leakage*, where unintended overlap between training and test data skews evaluation results, giving an inflated sense of model performance.

Beyond data, process-related issues also pose a threat to evaluations. *Confirmation bias* occurs when researchers selectively use data or metrics favouring their AI system, leading to overly optimistic assessments. Some systems may even be designed to exploit specific evaluation metrics or datasets – known as *gaming the system* – rather than demonstrating true generalization. Moreover, *incomplete comparisons* arise when only favourable metrics are highlighted, ignoring aspects where the system may underperform.

5.3.2 Possible Approaches

Based on our experience with the field and our discussion during the seminar, we identified the following initial strategies to address the issues presented in the previous section.

1. **Backtesting** involves training AI systems on data from one specific time period or geographic region and testing them on a different time period or region. For example, a model trained on European data may be evaluated on data from the U.S. to assess its adaptability across different contexts. Commonly used in causal discovery, this method evaluates robustness across temporal or spatial shifts.
2. **Lifelong Benchmarking.** Instead of relying solely on static benchmarks, this evaluation approach continuously updates benchmarks with new datasets and annotations while reusing previously validated models to test these new datasets. This dynamic approach ensures evaluations remain relevant over time.
3. **Reproducibility Testing** assesses whether an AI system's outcomes can be consistently replicated across different domains, datasets, or implementation approaches, highlighting the generalizability of the system.
4. **Sandboxing and Red Teaming**
   - **Sandboxing** implies deploying the AI system in a controlled environment, recording results and user interactions. It helps assess how the system's behaviour in different scenarios, records performance metrics, and gathers user feedback.
   - **Red Teaming** involves a dedicated team of experts to intentionally "break" the system, identifying weaknesses that standard evaluation might overlook and logging qualitative feedback.
5. **Coherence Testing** evaluates whether the AI system produces consistent and logically coherent results across similar queries, emphasizing internal reliability. Inconsistent results may indicate issues, such as inadequate training or overfitting to specific datasets/tasks.
6. **Evidence-Based Evaluation** measures the system's ability to provide well-supported, documented evidence for its outputs. The quality and amount of evidence are proxies for performance and key indicators of the system's reliability and trustworthiness.
7. **Explanation-Based Evaluation.** Explanations bridge an AI system's internal reasoning and human understanding, providing a basis for trust. High-quality explanations are clear, logical, and relevant, while poor-quality explanations can erode trust and perpetuate biases.

5.3.3 Connection to overarching topics

The seminar centred on four central aspects of KG-based AI systems: transparency, trust, accountability, and self-determination. One question is posed for each of these aspects; here, we reflect on these questions and illustrate how evaluation approaches can help address them.

Transparency *What is required to ensure that the data fueling and the inner workings of AI artefacts are transparent?*

Evidence-based Evaluation supports transparency by evaluating the capability of AI systems to provide well-supported, verifiable, and documented evidence for their outputs. The evaluation approach allows users to understand the basis of the system's decisions and verify the sources of information used. *Reproducibility Testing* also contributes to transparency by ensuring that results can be replicated across different domains and implementations.

Trust *What are the key requirements for an AI system to produce trustable results?*
Coherence Testing is essential for trust to demonstrate an AI system's ability to provide consistent and logically sound outputs. Other approaches can also improve the trust in an AI system, e.g., *Sandboxing and Red Teaming* by stress-testing the system under controlled conditions, identifying vulnerabilities, and ensuring robustness before full deployment, and *Lifelong Benchmarking* through dynamic updates of evaluation benchmark, making it relevant over time.

Accountability *How can AI be made accountable for its decision-making?*
Evidence-Based Evaluation enforces accountability by requiring AI systems to justify their decisions with verifiable data. Other approaches enhance accountability, e.g., *Explanation-Based Evaluation* that allows stakeholders to review how decisions are made, and *Red Teaming*, which helps identify weaknesses and biases in AI systems. Note that the latter delegates accountability to the stakeholders.

Self-Determination *How can users and citizens maintain self-determination when using or being the subject of KG-based AI systems?*
Transparency-focused methods such as *Evidence-Based Evaluation* empower users by giving them insight into how AI systems work, helping them to reason and make an informed decision whether and how to use AI system results. Other approaches also contribute to self-determination, e.g., *Lifelong Benchmarking*, which ensures models remain aligned with evolving user needs, societal norms, and ethical standards.

5.3.4 Conclusion

Traditional benchmarking approaches to AI evaluation have significant limitations, particularly for complex and dynamic domains. By integrating innovative strategies such as explanation-based evaluation, backtesting, and lifelong benchmarking, we can overcome these challenges and develop more robust, accountable neurosymbolic AI systems. As next steps, we see a need for:

- **Exploring evaluation strategies.** Implement and evaluate possible evaluation approaches, such as backtesting protocols for knowledge-graph-based AI systems and develop standardized protocols for red teaming and sandboxing.
- **Relating evaluation to architectures.** Investigation of evaluation strategies in the context of neurosymbolic AI design pattern architectures [3] to enhance their interpretability and robustness.

References

1 Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable AI. 10(1):147–159, 2022.
2 Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.
3 Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. Modular design patterns for hybrid learning and reasoning systems. *Appl. Intell.*, 51(9):6528–6546, 2021.

5.4 Trust and Accountability in Knowledge Graph-Based AI for Self Determination: Building World Models in Formal Representations using LLMs

José Manuel Gómez-Pérez (Expert.ai – Madrid, ES)
Marko Grobelnik ((Jozef Stefan Institute – Ljubljana, SI)
Ryutaro Ichise (Institute of Science Tokyo, JP)
Manolis Koubarakis (University of Athens, GR)
Heiko Paulheim (University of Mannheim, DE)
Daniel Schwabe (Rio de Janeiro, BR)

5.4.1 Motivation

Large Language Models (LLMs) have demonstrated impressive capabilities in understanding and generating natural language, including extracting structure and meaning from unstructured text. However, their internal world knowledge and reasoning processes remain largely opaque. In this working group, we investigated the hypothesis that it is possible to construct explicit, formal world models from the latent knowledge encoded in LLMs and textual sources – models that could serve as stable, inspectable, and verifiable representations of a domain.

This endeavor aligns with long-standing goals in artificial intelligence and knowledge engineering, particularly the knowledge acquisition bottleneck, which has historically hampered the development of large-scale, formalized knowledge bases[2]. Inspired by systems such as Cyc, which utilized microtheories to manage and modularize world knowledge, we explored whether modern LLMs could support a similar, but largely automated, process of world model construction – potentially generating formal representations in languages such as OWL, SWRL, or Prolog directly from text.

The core motivation rests on several anticipated benefits of this approach:

- **Improved reasoning:** Formal models constrain the solution space and enable more accurate, consistent, and explainable inferences than directly querying an LLM.
- **Explainability and transparency:** Representing knowledge symbolically allows inspection of both structure and content, addressing key concerns in trustworthy AI.
- **Reproducibility and auditability:** Formalized ontologies can be versioned, verified, and reused in a way that black-box LLM behavior cannot.
- **Human-AI collaboration:** LLMs can support domain experts and knowledge engineers in structuring, extending, and refining ontologies and rule-based systems.

As a working hypothesis, we posited that such LLM-bootstrapped world models – constructed through a combination of natural language interpretation and formal reasoning – could eventually lead to systems that are both expressive and interpretable, supporting multi-hop reasoning, comparative validation, and context-sensitive explanation.

This investigation was grounded in a concrete domain: AI regulation and compliance, specifically centered on the EU AI Act[1]. This domain combines rich natural language source texts, evolving legal definitions, and complex reasoning requirements, making it an ideal setting to evaluate the feasibility and value of constructing formal world models with LLM assistance.

5.4.2 Method and Experiments

Our approach combined classical knowledge engineering (KE) methodologies with the affordances of modern Large Language Models (LLMs), using the latter as active participants throughout the modeling pipeline. The LLMs used during our experimentation spanned across the spectrum of models in the Google Gemini and Open AI families. However, soon we settled for models with capabilities involving inference-time scaling[3], such as o1 and Gemini-2.5, which demonstrated to be better suited for complex reasoning tasks like this.

Rather than relying on LLMs solely as generators of content or answers, we treated them as collaborators in an iterative, human-in-the-loop process of constructing a formal world model from regulatory text. The process was inspired by standard KE lifecycles and followed an adapted knowledge engineering lifecycle, leveraging LLM capabilities at each stage:

1. **Domain and Scope Definition** We selected the EU AI Act as our primary source, focusing on its provisions for high-risk AI systems. This emerging legislation offers a structured yet complex body of natural language, with deep implications for the classification, deployment, and oversight of AI systems. This domain was chosen based on group expertise, real-world relevance, and the presence of rich, non-trivial natural language source material. We narrowed our focus to Title III and Annex III of the AI Act, which provide definitional content, classification rules, technical requirements, and compliance procedures. The textual structure of the act itself is quite complex. We first asked the LLM to analyze its structure and identify the portions relevant to high-risk systems and compliance, using these sections to drive world model extraction.
2. **Competency Question Generation** To frame the modeling effort, we generated a set of competency questions – queries that the resulting world model should be able to answer. These ranged from high-level regulatory assessments (e.g., "Is this product considered high-risk under the AI Act?") to accountability and governance inquiries (e.g., "Who is responsible for ensuring compliance?"). We created approximately 25 such questions, some crafted manually, and others generated by prompting LLMs such as ChatGPT or Gemini with high-level modeling intents. These questions helped clarify modeling scope and purpose, and provided a foundation for validating the emerging models.
3. **Requirements Gathering and Ontology Design** Using LLMs, we identified relevant regulatory texts, conceptual elements, and ontologies. This included mining the AI Act for definitions, roles, obligations, and processes, as well as exploring existing semantic vocabularies (e.g., LegalRuleML, FOAF, PROV-O) for potential reuse. Prompts were used to generate OWL ontologies in Turtle syntax, as well as Prolog-style FOL rules. LLMs effectively recommended languages (e.g., OWL, SWRL, Prolog), modeling strategies, and external ontologies, showcasing multi-representational flexibility by shifting between different formalisms as needed.
4. **Ontology Construction and Extension** Initial top-down ontologies were produced by feeding full or partial text of the AI Act to LLMs. Later iterations involved incremental refinements based on new examples or competency questions. We explored structuring models into reusable fragments, akin to microtheories, to enhance modularity. However, incremental modeling proved brittle: LLMs struggled to maintain consistency across iterations, often introducing duplication, inconsistency, or reference drift. Prompt sensitivity remained high, requiring careful tuning.
5. **Instance-Level Annotation** We selected Waymo One, an autonomous driving platform, as a representative AI system. LLMs were used to annotate product descriptions with formal instance data. Prompts requested annotation using previously generated

ontological classes. This exercise revealed key challenges: LLMs frequently introduced new namespaces or redundant concepts instead of reusing prior structure, underscoring limitations in consistency and reuse.

6. **Validation and Demonstration** Although full reasoning tests were out of scope during the seminar, we emphasized competency-query-based evaluation – ensuring that the formal model could correctly classify, explain, or reject assertions about specific AI systems based on regulatory criteria. Initial validations were performed using both OWL reasoners and Prolog engines. These early-stage demonstrations illustrated how symbolic reasoning could support competency question answering, such as "Why is this AI system considered high-risk?" or "Who holds compliance responsibility in this scenario?"
7. **Reflection and Iteration** Throughout the process, we embraced an iterative, bi-directional workflow – moving from text to formalism and back again, with LLMs supporting each translation. This round-trip modeling paradigm allowed for rapid prototyping and exploration of design alternatives. LLMs played multiple roles: advisor, translator, editor, explainer, and generator. Their ability to contextualize text, propose structured models, and convert between representational forms enabled fluid movement between informal and formal levels of abstraction.

5.4.3 Lessons Learned

Our experiments yielded a range of insights into the strengths, weaknesses, and emerging design patterns associated with using LLMs for world model construction. These lessons fall broadly into two categories: capabilities and affordances, and challenges and limitations.

5.4.3.1 Capabilities of LLMs in Knowledge Engineering

LLMs proved valuable collaborators in a variety of modeling tasks, particularly in the early phases of formalization:

- **Bootstrapping structured representations:** LLMs effectively proposed taxonomies, relations, and axioms from unstructured legal text. With the right prompting, they could output OWL in Turtle syntax or Prolog-style logic programs.
- **Generating competency questions:** When guided appropriately, LLMs generated high-quality, domain-relevant competency questions, helping to clarify modeling scope and purpose.
- **Advising on tools and methodologies:** LLMs could recommend languages (e.g., OWL, SWRL, Prolog), suggest modeling strategies, and identify relevant external ontologies.
- **Multi-representational flexibility:** LLMs easily shifted between different formalisms, such as transforming an OWL ontology into SWRL rules, or converting plain text into logic-based representations.
- **Prototyping and iteration:** LLMs supported rapid prototyping of models and allowed quick exploration of design alternatives – essential for an exploratory setting like a Dagstuhl Seminar.

A key observation was that LLMs were most effective when deployed in a hybrid approach, mixing: **top-down** modeling from source texts and domain concepts, and **bottom-up** enrichment via instance-level annotations and use-case reasoning. This interplay allowed for richer and more grounded models – though not without difficulty.

5.4.3.2 Challenges and Limitations

Despite their impressive capabilities, LLMs exhibited several recurring limitations that hindered more robust modeling workflows:

- **Incremental modeling is brittle:** LLMs struggled to maintain consistency across iterations. Changes to a previously generated ontology often led to duplication, inconsistency, or regression.
- **Reference drift:** LLMs frequently failed to reuse previously defined classes, properties, or namespaces, even when explicitly instructed to do so.
- **Prompt sensitivity:** Small changes in prompt structure often led to significantly different outputs. Prompt engineering required careful tuning and could not be reliably abstracted.
- **Implicit vs. explicit knowledge:** When asked to extract axioms from text, LLMs frequently filled in gaps with inferred or assumed knowledge, blurring the boundary between source-derived content and background knowledge.
- **Provenance and traceability:** It was difficult to track which parts of the generated model were based on the original source text versus LLM inference or hallucination.
- **Toolchain fragility:** Integration with formal modeling tools (e.g., OWL editors, Prolog reasoners) exposed limitations in both syntax fidelity and model completeness.

These challenges underscored the need for better support for human-in-the-loop verification, robust referencing, and systematic iteration when using LLMs in formal modeling tasks.

5.4.4 Open Questions and Future Work

While our initial experiments confirmed the potential of LLMs to assist in formal world model construction, they also raised a number of open questions and future research directions. These span technical, methodological, and conceptual dimensions.

5.4.4.1 Methodological Open Questions

- **How should human–AI collaboration be structured in formal modeling?** The process we followed was ad hoc but promising. More systematic methodologies are needed to orchestrate interactions between domain experts, knowledge engineers, and LLMs – particularly in maintaining consistency across iterations.
- **What does "correctness" mean in this context?** Unlike traditional logic programs or ontologies, LLM-generated models may reflect probabilistic or context-dependent interpretations. Determining when a world model is "good enough" or "trustworthy" remains an open question.
- **What role should competency questions play in validation?** Competency questions helped frame the modeling task, but their use as a systematic evaluation mechanism is underdeveloped. Future work could define benchmarks or test suites to assess model coverage and consistency.

5.4.4.2 Technical Challenges and Research Directions

- **Consistency and Reuse** A key technical challenge is ensuring that LLMs consistently reuse previously defined structures – classes, properties, namespaces – across sessions and modeling phases. This requires better prompt design, memory management, and possibly external constraint injection.

- **Iterative, verifiable modeling** Formal models need to evolve incrementally without introducing contradictions or drift. Future research could explore interfaces where LLMs propose edits that are checked by reasoners or validated against typed assertions.
- **Toolchain integration** Bridging the gap between LLMs and formal reasoning systems remains a challenge. Improved APIs, modeling environments, and reasoning-aware LLM prompts could help operationalize neurosymbolic modeling pipelines.
- **Round-trip modeling and explanation** One of the most exciting possibilities is the idea of a round-trip loop between symbolic models and LLMs: models extracted from LLMs are refined by humans and used in turn to improve LLM outputs or explanations. Supporting this requires mechanisms for traceability, grounding, and goal-sensitive reasoning.

5.4.4.3 Broader Implications

- **From world models to world views** If LLMs can generate multiple, diverging models from the same text (as our experiments suggest), this opens up the possibility of comparing different stakeholder perspectives – e.g., between a regulator, a provider, and an affected user. This introduces new opportunities for accountability, argumentation, and normative reasoning.
- **Explainability as emergent behavior** Formal models on their own do not constitute explanations – but they provide the scaffolding for tailored, stakeholder-specific narratives. Future work could explore how formal world models and LLM-based natural language generation can jointly support explainable AI in high-stakes domains like regulation.

References

1 European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024, 2024.
2 Cogan Shimizu and Pascal Hitzler. Accelerating knowledge graph and ontology engineering with large language models. *Journal of Web Semantics*, 85:100862, 2025.
3 Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv*, arXiv:2408.03314, 2024.

5.5 Knowledge Graph Ecosystems

Sandra Geisler (RWTH Aachen, DE)
James A. Hendler (Rensselaer Polytechnic Institute – Troy, US)
Philipp D. Rohde (TIB – Hannover, DE)
Aisling Third (The Open University – Milton Keynes, GB)
Maria-Esther Vidal (TIB – Hannover, DE)

5.5.1 Introduction

In this working group, we based our initial discussions on the concepts of Knowledge Graph Ecosystems (KGEs) inspired by the paper by Geisler et al.[1] In the paper, knowledge graph ecosystems describe the context and crucial aspects impacting the creation, updates, and usage of knowledge graphs by a sextuple. This includes the data sources populating the KG,

the ontologies to enrich and structure the data in the KG, mappings between the ontology and the data, constraints, which enable consistency and quality checks, the KG which is "living" in the KGE and subject to changes and usage, and finally a log, tracing the operations executed on the KG and corresponding events. Further, the paper delineates life cycles and life cycle steps which structure and order the operations on the KG. A life cycle step can be, e.g., the creation or the update of a KG, or a query or analysis on the data of the KG. Formally, it is defined also as a sextuple with a service executing the step, actors and roles involved in the step, as well as requirements, constraints, and needs which impact the life cycle step. While this work is a crucial step towards the formalization and therefore automatization and verifiability of operations on and around KGs, it only targets individual KGs. However, today many applications require sharing data, i.e., the querying and analysis of multiple distributed KGs across organizations, domains, and even countries. Hence, we need to not only take into account contexts and life cycles of single KGEs, but also networks or federations of KGEs.

In exploring the concepts of federated and decentralized Knowledge Graph Ecosystems (KGEs), this work inherently ties into the foundational themes of trust, accountability, and autonomy discussed in [2]. Note that each of these themes are very broad concepts with significant variation in their concrete application; for a fully flexible model of KGEs, it would be a mistake to attempt to narrow trust, accountability, and autonomy down to very specific or technical definitions only covering a subset of their broader meanings. We rather leave the framework open to accommodate different concrete understandings of these terms in different scenarios or use cases. Trust in federated KGEs is established through shared ontologies and robust data verification processes, similar to the way blockchain's transparency and immutability enhance trust by ensuring the traceability and provenance of data. This aligns with decentralized KGEs where trust mechanisms, like reputation systems, are vital for ensuring knowledge integrity. Accountability is addressed through verifiable contributions and collaborative data management in federated networks, paralleling the mechanisms in [2], such as compliance-checking and provenance tracking that uphold data integrity through blockchain technology. Autonomy is a defining feature of decentralized KGEs, where stakeholders maintain control over their data and operations, consistent with the use of decentralized infrastructures in [2], which empower individuals through self-sovereign identities and personalized data management. By integrating these technologies, both papers emphasize the potential to advance KGEs in ways that are reliable, user-centric, and conducive to collaborative innovation.

In the following, we will present the examples for such networks and federations of KGs discussed in the seminar. Based on these examples, we will delineate the different types of KGEs we derived from the examples and discuss challenges and requirements for the different types. These challenges and requirements led to a reformulation and extension of the definitions for KGEs and the corresponding life cycles, which we will sketch subsequently. Finally, we will give an outlook on future work sparked by the results of the seminar.

5.5.1.1 Use Case 1: Digital Product Passports

Transparency of product constitution, production processes, and supply chains, especially in the light of sustainability, energy consumption, and circular economies, motivates the need for a Digital Product Passport (DPP). A DPP serves as a comprehensive digital record that comprises data about a product's components and life cycle, including its materials, manufacturing processes, usage, and recycling potential. DPPs are promising transparency and traceability, ensuring that all stakeholders – from manufacturers to consumers – can

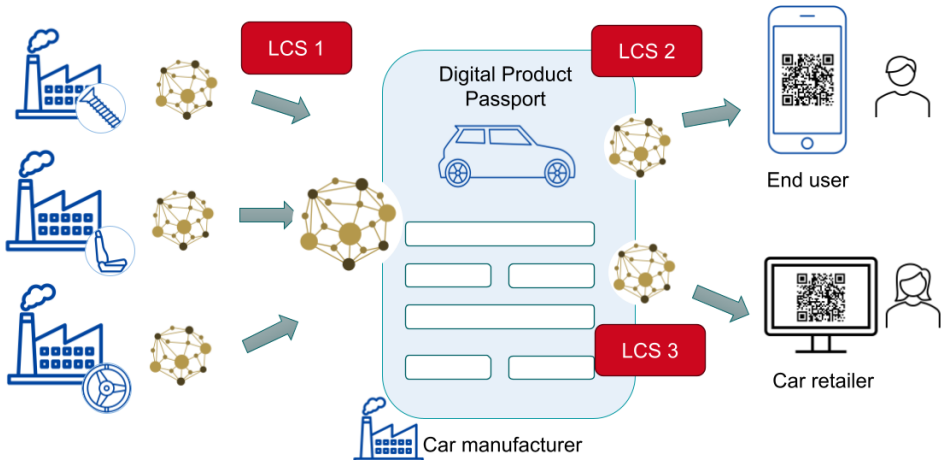


Figure 3 A KGE of a Digital Product Passport.

make informed decisions while fostering accountability. Additionally, a DPP can facilitate compliance with regulatory requirements, improve waste management, and support efficient product recalls or repairs. However, the data for the DPP needs to be retrieved from the involved stakeholders. Suppliers, logistics, as well as manufacturers, can provide information regarding a product or parts of it. Furthermore, usage behavior, repairs, and other events related to the product user and impacting the product quality and integrity can also be part of a DPP.

Figure 3 shows the example of a DPP infrastructure for a car. Suppliers of screws, seats, or steering wheels all may maintain data about the part they are providing. It is assumed that product data for the DPP is stored in KGEs located at the suppliers. To enable a DPP, the suppliers extract and provide a part of this data to the car manufacturer who in our example is assumed to maintain the DPP of the car, i.e., integrate the data from the suppliers into a bigger KGE. We assume further that different stakeholders may have different views on the KGE of the product maintained by the manufacturer. Thereby, subsets in the form of views can be extracted from the KG of the DPP and are provided to consuming applications, such as mobile applications for end users or web applications for retailers. In this scenario, the stakeholders follow a common goal, namely to provide a DPP. In terms of trust, the car manufacturer trusts that the suppliers provide the DPP information of their part, i.e., they are accountable for the provision of correct, accurate, and complete data. They do not have a high autonomy as they are dependent on the car manufacturer as their customer and follow its rules.

5.5.1.2 Use Case 2: A Dagstuhl Seminar

Opposed to the example of a DPP, in a Dagstuhl Seminar multiple stakeholders (the participants and organizers) attend with one or more individual goals for the overall seminar, but do not necessarily share a common goal. They all have their own "internal KGE" and may update their knowledge in interaction with other participants. Further, they build smaller working groups, where the group follows common group goals and at the same time can contribute to the overall goal of the seminar. Hence, there are multiple KGEs (one for each participant, one for a group, and one for the seminar), which have their own life cycles, "interact" with each other and exchange knowledge between them. Between the participants,

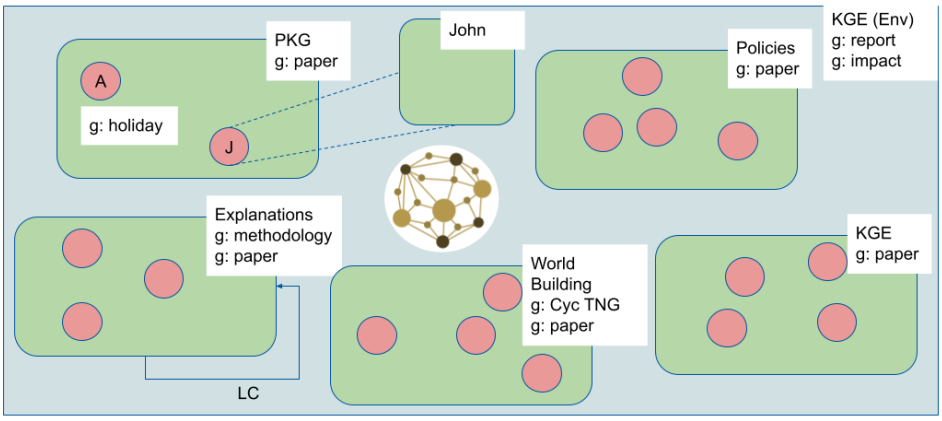


Figure 4 The KGEs of a Dagstuhl Seminar.

trust needs to be established as a basis to update their KGEs in the interaction. There is a very high autonomy as no participant is dependent on another and in consequence also no accountability.

5.5.2 Types of KGE Networks

We define a KGE network as a set of KGEs, which are interconnected by relationships.

We have considered each of the examples above in terms of how they demonstrate different assumptions about the three pillars of trust, accountability, and autonomy, and how they relate to each other. By configuring KGE networks with these assumptions, we can see that network phenomena such as *federation*, decentralization, and so on, emerge. In the DPP example, low autonomy of stakeholders, represented by the common specification set by the car manufacturer, and trust ensured via contractual accountability, looks very like a *federated* scenario; a Dagstuhl Seminar's high autonomy of participants, with low accountability and high trust, resembles more a *decentralized* scenario.

A relationship is established between two KGEs to enable knowledge exchange between them. Based on the observations from the above two and also other examples, we distinguish the following two types of KGE networks:

1. **Federated KGE Networks:** In a federated KGE network all stakeholders have a KGE and share a common goal towards which they exchange knowledge.
2. **Decentralized KGE Networks:** In this type of networks stakeholders also have a KGE, but may have one or more individual goals. Additionally, they can also share common goals with one or more other stakeholders.

These definitions are in terms of entities formally defined in the KGE framework of [1]: goals, knowledge graphs, etc. We show below how these can nonetheless provide models of these trust, accountability, and autonomy scenarios. This provides an insight into how to model scenarios involving the foundational pillars of [2] without overcommitting to narrow definitions of them, and we argue that this flexibility coupled with the representation formalism of KGEs is essential for AI ecosystems which work for such fundamentally human concepts.

5.5.3 Requirements

Understanding the requirements inherent in different use cases of Knowledge Graph Ecosystems (KGEs) provides insight into the diversity of challenges and solutions necessary to support various applications. This section describes the requirements for the two primary types of KGE networks – Federated and Decentralized – each characterized by unique aspects of collaboration, autonomy, trust, and lifecycle management.

5.5.3.1 Federated KGE Networks

The case of Digital Product Passports (DPP) serves as an archetype for federated KGEs where stakeholders share a collective objective and are inherently dependent on each other's data integrity and contributions. The requirements for such setups include:

1. **Shared Semantics and Ontologies:** Stakeholders must align on a common ontology to ensure interoperability across KGEs. This includes shared vocabularies and data models that describe products consistently.
2. **Data Integration and Consistency:** The integration of data from multiple suppliers entails rigorous consistency checks and validation mechanisms to maintain data integrity across the ecosystem.
3. **Accountability and Verifiable Contributions:** Federated systems necessitate mechanisms for traceable contributions to ensure that each actor can be held accountable for the data they provide, thus fostering trust in the overall system.
4. **Role-Based Access and Governance:** Clearly defined (e.g., suppliers, manufacturers, regulators) must be supported, with correspondingly specific access rights and responsibilities. Governance models ensure appropriate access and usage of the ecosystem by different stakeholders.
5. **Lifecycle Synchronization:** Operations across constituent KGEs must be coordinated, especially for updates and validations, to maintain a consistent state and ensure collaborative workflows proceed smoothly.

5.5.3.2 Decentralized KGE Networks

In contrast, decentralized KGE networks as exemplified by collaborative scenarios like scientific research or Dagstuhl Seminars, emphasize autonomy and less rigid interactions. The requirements for decentralized structures include:

1. **Alignment Mechanisms:** Decentralized networks require lightweight ontology alignment methods to facilitate knowledge exchange without enforcing uniformity, thus respecting participants' autonomy.
2. **Trust-Based Interactions:** Establishing trust through provenance and reputation mechanisms is critical in decentralized setups where actors operate independently but collaborate based on shared interests or goals.
3. **Asynchronous and Independent Lifecycles:** Supports for asynchronous updates and independent lifecycle management are needed, allowing participants to operate under varied goals and timelines without disrupting mutual interactions.
4. **Autonomy Preserving Policies:** Each participant should have the autonomy to apply local policies and operate independently, with knowledge exchange only occurring in mutually beneficial circumstances.
5. **Conflict Resolution and Negotiation Frameworks:** Mechanisms to resolve conflicts or inconsistencies are essential where differing goals or data interpretations may arise from the diverse participants.

5.5.3.3 Challenges and Opportunities

The interplay between these requirements in both federated and decentralized networks highlights several overarching challenges:

1. **Interoperability:** Striking a balance between shared understanding and local autonomy requires robust semantic alignment techniques applicable across diverse contexts.
2. **Trust and Accountability:** Crafting ecosystems where trust and accountability are clear yet flexible enough to accommodate decentralized decision-making and governance remains a pivotal challenge.
3. **Efficiency and Scalability:** Efficient data management, processing, and query execution across distributed KGEs need to support scalability while preserving data quality and provenance.

Addressing these requirements not only involves technical innovation in areas like ontology matching, knowledge integration, and lifecycle management but also necessitates careful consideration of legal, ethical, and organizational factors to foster successful KGE deployments. As such, continued research in synchronizing lifecycles, implementing robust alignment frameworks, and enhancing-based governance will pave the way for more adaptable and scalable KGE networks.

5.5.4 Extension of the KGE Concept

The original concept of a Knowledge Graph Ecosystem (KGE), as introduced by Geisler et al. [1], models the fundamental operational components required to manage the creation, evolution, and analysis of a knowledge graph. Formally, a KGE is defined as a sextuple: $KGE = (D, O, M, DC, KG, L)$, where:

- D is a set of **data sources**, each with schema $\theta(ds)$ and instances $\alpha(ds)$.
- O is the **ontology**, a logical theory describing the domain using a structured vocabulary.
- M is a set of **mappings** linking D to O through semantic assertions.
- DC is a set of **domain constraints** ensuring data consistency and quality.
- KG is the resulting **knowledge graph**, built from D using M under O.
- L is a **log** of lifecycle steps, tracking changes and ensuring traceability.

Limitations. While this formalization provides a solid foundation for the structured management of a single evolving knowledge graph, it assumes:

1. A single, self-contained knowledge graph (KG) under centralized control.
2. A fixed and isolated lifecycle of operations for one ecosystem context.
3. No explicit support for collaboration or interactions across multiple independent KGEs.

However, real-world applications increasingly rely on *networks of interconnected KGEs* operated by diverse stakeholders. These scenarios require knowledge to be collaboratively constructed, exchanged, and reused across institutional, geographical, or organizational boundaries. Consider the previously defined use cases:

- **Digital Product Passports (DPPs)**: A manufacturer aggregates data from multiple suppliers, each maintaining their own KGE to describe components or materials. The manufacturer composes a global KGE that integrates these subgraphs. All actors are accountable for their data and contribute toward a shared objective.
- **Collaborative Scientific Research**: Researchers maintain local KGEs to organize data and hypotheses. As part of a collaborative research initiative, they align concepts and exchange selected data views, but retain autonomy in their lifecycle and modeling.

To enable such scenarios, we extend the original definition of KGE by introducing a recursive structure where ecosystems can consist of multiple interrelated KGEs – each with its own lifecycle, goals, and context.

Extended Definition. A Knowledge Graph Ecosystem is now defined as:
$KGE = (KGM, \{KGE_1, \ldots, KGE_n\}, LC, G, C)$, where:
- $KGM = (D, O, M, DC, KG, L)$ is the underlying **knowledge graph model**.
- $\{KGE_1, \ldots, KGE_n\}$ is a (possibly empty) set of **nested or interacting KGEs**.
- LC is a **lifecycle**, formally modeled as a partially ordered structure of lifecycle steps, each representing operations or analyses on the KG.
- G is a set of **goals** pursued by the ecosystem.
- C is the **context**, capturing domain-specific and regulatory constraints or assumptions.

This extended definition allows the modeling of composite or hierarchical KGEs, enabling both *federated* and *decentralized* structures.

Federated KGEs. A KGE is **federated** if it aggregates multiple KGEs that share:
- **Common goals and context**: $G \subseteq \bigcap_{i=1}^{n} G_i$ and $C \subseteq \bigcap_{i=1}^{n} C_i$.
- **Shared concepts**: $O \models \bigcup_{i=1}^{n} O_i$ and $KG \subseteq \bigcup_{i=1}^{n} KG_i$.
- **Shared data**: $D \subseteq \bigcup_{i=1}^{n} D_i$.

This setup supports alignment and accountability among actors collaborating toward a unified goal, such as regulatory compliance or industry standards.

Use Case 1 (DPP) is an example of a federated KGE, where suppliers and manufacturers contribute to a shared infrastructure, follow a common goal, and are held accountable for the correctness of their data.

Decentralized KGEs. A KGE is **decentralized** if it includes autonomous KGEs that:
- Maintain **individual goals and contexts**, i.e., $G \cap \bigcup G_i \neq \emptyset$ or $C \cap \bigcup C_i \neq \emptyset$.
- Communicate via **alignments**: for each pair (i, j), there exists an ontology O_{ij} such that $O_i \models_a O_{ij}$ and $O_j \models_a O_{ij}$, where $\models_a$ denotes entailment up to alignment.
- Possibly share some **data elements**.

Such configurations reflect looser, trust-based networks – e.g., research collaborations or international knowledge exchanges – where autonomy is preserved, and accountability is local. **Use Case 2 (Dagstuhl Seminar)** exemplifies this structure, with researchers individually maintaining their KGEs while participating in collaborative groups with overlapping knowledge and trust-based data sharing.

Summary. Extending the KGE model from single knowledge graphs to federated and decentralized ecosystems enables:
- Formal modeling of complex, distributed knowledge infrastructures.
- Explicit lifecycle tracking across ecosystem components.
- Representation of autonomy, trust, and alignment across diverse actors.

These extensions are essential to support knowledge-centric systems that operate across organizational, geographic, and technical boundaries.

5.5.5 Challenges and Future Work

Extending the KGE concept to support federated and decentralized ecosystems introduces new research challenges that go beyond those of traditional or standalone knowledge graphs. These challenges stem from the need to enable collaboration, preserve autonomy, ensure accountability, and foster trust across heterogeneous, interlinked systems.

1. Semantic Alignment in Heterogeneous Ecosystems. While classical KGEs already face challenges in integrating heterogeneous data, these are further exacerbated in federated and decentralized settings. Each participating KGE may use distinct ontologies, vocabularies, and constraints, making alignment a prerequisite for knowledge exchange. Federated settings assume shared semantics; decentralized ones require ontology alignments that preserve local autonomy while supporting partial interoperability. Ontology matching, cross-KGE mappings, and alignment reasoning are required to maintain interoperability and trust.

2. Trust and Accountability in Distributed Architectures. As KGEs become interlinked, accountability must be clearly defined across organizational and jurisdictional boundaries. Federated KGEs require all participants to commit to shared goals and contexts, making it possible to define collective accountability [2]. Decentralized KGEs, by contrast, emphasize autonomy – participants operate independently but must still be trusted sources. Establishing trust involves formalizing provenance, defining verifiable contributions, and enabling transparency in lifecycle actions. Ecosystem-wide logging mechanisms and policy-compliant data usage protocols are needed to support accountability and traceable decisions.

3. Modeling Autonomy and Interactions. In decentralized KGEs, each participant may pursue different goals or operate under distinct regulatory or ethical frameworks. Maintaining autonomy requires the ability to define local policies, ontologies, and lifecycle models – while still enabling interaction through lightweight semantics and alignments. Ensuring that knowledge exchange does not infringe on local autonomy calls for formal contracts, soft alignments, and partial knowledge views. Such mechanisms allow for trust-based collaboration while upholding the principle of self-determination [2].

4. Lifecycle Across KGEs. Federated and decentralized KGEs introduce the challenge of coordinating lifecycles across independently evolving components. In federated settings, lifecycle steps (e.g., updates, validations) must be synchronized to ensure consistent outcomes. In decentralized settings, looser coordination must support asynchronous updates and versioning. Future work should explore distributed lifecycle management models, including temporal consistency, event-driven propagation, and local override mechanisms.

5. KGE Validation and Provenance. Validation across KGEs requires understanding how constraints defined in one ecosystem apply to others. This is particularly challenging when KGE components evolve independently or when only partial views are exchanged. Provenance models must capture not only source alignment but also how knowledge was transformed or aligned across systems. Explainability – especially for outputs from KG-based AI systems – requires tracing decisions back to source KGEs and validating their integrity.

6. Explainability and Repeatability in KG-Based AI. As KGEs serve as the backbone for AI systems, the ability to explain inferences and reproduce results is crucial for building trust. This requires detailed provenance, clear alignment semantics, and interpretable reasoning paths – especially when hybrid approaches (e.g., involving LLMs) are used. Repeatability in decentralized settings must account for potential disappearance or evolution of external data sources. Mitigation mechanisms, such as fallback reasoning strategies or versioned snapshots, are essential to uphold the trustworthiness and reliability of AI outcomes.

7. Defining Role-Based Trust and Governance Models. With multiple stakeholders, each with distinct roles (e.g., data provider, knowledge builder, auditor, consumer), KGEs must support differentiated views, permissions, and responsibilities. Role-specific governance models and access policies must ensure that users interact with the ecosystem in ways that are transparent, justified, and traceable. This includes defining what actions are permissible, who is accountable for data changes, and how conflicts or inconsistencies are resolved.

8. Modeling Roles, Personas, and Task-Specific Interactions. KGEs serve a diverse set of stakeholders – including data stewards, domain experts, developers, auditors, and end users – each with distinct objectives, capabilities, and responsibilities. The extension to federated and decentralized KGEs further amplifies this diversity, requiring systems to explicitly model and support *role-specific interactions*. These roles influence how users contribute to, consume from, or govern the KGE. For example, a data provider must ensure schema and content quality, while a consumer needs trustable and explainable insights. Supporting personas involves:

- Designing user interfaces and APIs tailored to task-specific workflows.
- Implementing role-based access control and provenance-based accountability.
- Supporting transparency through lifecycle-aware logs and explainable outputs aligned with each persona's mental model.

Future work must focus on formalizing persona definitions, mapping tasks to lifecycle stages, and capturing the responsibilities of each actor. These models are essential to foster trust and usability in multi-actor KGEs, and to ensure that autonomy and accountability are respected across stakeholder boundaries.

Summary. Federated and decentralized extensions of KGEs offer a powerful abstraction for supporting complex, multi-actor knowledge infrastructures. However, realizing these systems at scale requires addressing a new set of challenges around interoperability, lifecycle synchronization, role-specific trust, and accountability. Future research must focus on:

- *Trust*: Establishing mechanisms for verifiable contributions, alignment justifications, and explainable inferences.
- *Accountability*: Formalizing roles, logs, and lifecycle provenance to assign responsibility and enable redress.
- *Autonomy*: Supporting local lifecycles, policies, and ontologies while enabling coherent interactions across KGEs.

These directions are essential to align the evolution of KGEs with principles outlined in the proposed EU AI Act [11], including transparency, reliability, and respect for autonomy.

References

1 Sandra Geisler, Cinzia Cappiello, Irene Celino, David Chaves-Fraga, Anastasia Dimou, Ana Iglesias-Molina, Maurizio Lenzerini, Anisa Rula, Dylan Van Assche, Sascha Welten, et al. From genesis to maturity: Managing knowledge graph ecosystems through life cycles. *Proceedings of the VLDB Endowment*, 2025.

2 Luis-Daniel Ibáñez, John Domingue, Sabrina Kirrane, Oshani Seneviratne, Aisling Third, and Maria-Esther Vidal. Trust, accountability, and autonomy in knowledge graph-based AI for self-determination. *TGDK*, 1(1):9:1–9:32, 2023.

[11] https://artificialintelligenceact.eu/the-act/

6 Conclusions

John Domingue (The Open University – Milton Keynes, GB, john.domingue@open.ac.uk)
Luis-Daniel Ibáñez (University of Southampton, GB, L.D.Ibanez@soton.ac.uk)
Sabrina Kirrane (Vienna University of Economics and Business, AT, sabrina.kirrane@wu.ac.at)
Maria-Esther Vidal (TIB – Hannover, DE, vidal@l3s.de)

In conclusion, this seminar stands as a pivotal gathering that convened researchers and industry partners from diverse backgrounds. Together, we explored the complexities, challenges, and advancements inherent in managing and leveraging knowledge graphs within real-world contexts. Spanning from technical considerations to social dimensions, we identified essential requirements, imperatives, and actionable strategies necessary to foster the development of a new generation of knowledge graph ecosystems.

Given the advent of generative AI and its demonstrated benefits when integrated with intricate data structures such as knowledge graphs, ensuring readiness across all facets of knowledge graph implementation is paramount. The convergence of knowledge graphs with emerging technologies presents novel avenues for advancing knowledge representation, reasoning, and applications. Our discussions underscored the significance of robust quality assessment mechanisms and stressed the importance of integrating human expertise and feedback loops throughout the knowledge graph lifecycle. From an educational standpoint, it is imperative for experts to disseminate their knowledge through educational programs tailored to different levels of learning and professional training. However, standardizing competencies across all levels is essential to ensure a uniform understanding of fundamental concepts among potential knowledge graph practitioners.

Participants

- Sören Auer
TIB – Hannover, DE
- Piero A. Bonatti
University of Naples, IT
- Irene Celino
CEFRIEL – Milan, IT
- Andrea Cimmino
Polytechnic University of Madrid, ES
- Michael Cochez
VU Amsterdam, NL
- John Domingue
The Open University – Milton Keynes, GB
- Michel Dumontier
Maastricht University, NL
- Fajar Ekaputra
Vienna University of Economics and Business, AT
- Nicoletta Fornara
University of Lugano, CH
- Sandra Geisler
RWTH Aachen, DE
- Anna Lisa Gentile
IBM Almaden Center – San Jose, US
- José Manuel Gómez-Pérez
Expert.ai – Madrid, ES
- Marko Grobelnik
Jozef Stefan Institute – Ljubljana, SI
- Paul Groth
University of Amsterdam, NL
- Peter Haase
Metaphacts – Walldorf, DE
- Andreas Harth
Fraunhofer IIS – Nürnberg, DE
- Olaf Hartig
Linköping University, SE
- James A. Hendler
Rensselaer Polytechnic Institute – Troy, US
- Aidan Hogan
University of Chile – Santiago de Chile, CL
- Katja Hose
TU Wien, AT
- Luis-Daniel Ibáñez
University of Southampton, GB
- Ryutaro Ichise
Institute of Science Tokyo, JP
- Ernesto Jiménez-Ruiz
City St George's, University of London, GB
- Timotheus Kampik
SAP Berlin, DE & Umeå University, SE
- George Konstantinidis
University of Southampton, GB
- Manolis Koubarakis
University of Athens, GR
- Deborah L. McGuinness
Rensselaer Polytechnic Institute – Troy, US
- Julian Padget
University of Bath, GB
- Monica Palmirani
University of Bologna, IT
- Heiko Paulheim
University of Mannheim, DE
- Philipp D. Rohde
TIB – Hannover, DE
- Daniel Schwabe
Rio de Janeiro, BR
- Oshani Seneviratne
Rensselaer Polytechnic Institute – Troy, US
- Chang Sun
Maastricht University, NL
- Aisling Third
The Open University – Milton Keynes, GB
- Ruben Verborgh
Ghent University, BE
- Maria-Esther Vidal
TIB – Hannover, DE
- Jesse Wright
Open Data Institute – London, GB

Report from Dagstuhl Seminar 25052

From Research to Certification with Data-Driven Medical Decision Support Systems

Raul Santos-Rodriguez*[1], Kacper Sokol*[2], Julia E. Vogt*[3], and Sven Wellmann*[4]

1 **University of Bristol, GB.** enrsr@bristol.ac.uk
2 **ETH Zürich, CH.** kacper.sokol@inf.ethz.ch
3 **ETH Zürich, CH.** julia.vogt@inf.ethz.ch
4 **Universität Regensburg, DE.** sven.wellmann@barmherzige-regensburg.de

Abstract

This report outlines the programme and outcomes of Dagstuhl Seminar 25052 "From Research to Certification with Data-Driven Medical Decision Support Systems". Our seminar addressed the complex challenges of transferring artificial intelligence systems from research labs into real-world clinical practice. Bringing together clinicians, researchers and industry stakeholders, it explored the potential and pitfalls of deploying data-driven models in healthcare, highlighting the need for rigorous evaluation, human-centred design and responsible innovation. Key discussions included regulatory hurdles, reproducibility issues, interpretability and human–machine collaboration. Group sessions focused on evaluation frameworks and human factors in medical artificial intelligence system design. The seminar laid the foundation for a collaborative research agenda aimed at safe, effective and ethical integration of data-driven predictive models into real-life clinical workflows.

Seminar January 26–31, 2025 – https://www.dagstuhl.de/25052
2012 ACM Subject Classification Human-centered computing; Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning
Keywords and phrases artificial intelligence, clinical practice, decision support systems, digital healthcare, machine learning
Digital Object Identifier 10.4230/DagRep.15.1.201

1 Executive Summary

Kacper Sokol (ETH Zürich, CH)
Raul Santos-Rodriguez (University of Bristol, GB)
Julia E. Vogt (ETH Zürich, CH)
Sven Wellmann (Universität Regensburg, DE)

Seminar Vision

Artificial intelligence has made tremendous strides across many spheres of life, however deploying this technology in safety critical domains remains challenging. This Dagstuhl Seminar focuses on clinical practice where data-driven models can streamline the work of healthcare professionals and democratise access to personalised medicine, thus have lasting positive impact on society, but also where deploying such tools without adequate foresight and safeguards can be perilous. This duality – anticipated benefits that may come along

* Editor / Organizer

From Research to Certification with Data-Driven Medical Decision Support Systems, *Dagstuhl Reports*, Vol. 15, Issue 1, pp. 201–220
Editors: Raul Santos-Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

with unintended consequences – requires new technologies to be thoroughly vetted, e.g., with clinical trials and medical certification processes, before they can be deployed to avoid any harmful fallout. However, fulfilling such regulatory requirements is a lengthy and complex process plagued with many challenges, hence while prototype systems are becoming increasingly ubiquitous, they often remain indefinitely designated as research tools that can be used exclusively for research purposes. Their lacklustre adoption is compounded by pervasive reproducibility issues; history of unsafe systems being deployed prematurely; scarce data that are inherently private, difficult to collect or share, and often riddled with numerous biases; and prevalence of automation promises that never come to fruition. Such hurdles result in healthcare remaining one of the least digitised spheres of life.

A different contributing factor is predictive systems often being misconstrued as autonomous rather than social and relational, which is manifested in a counterproductive drive to match or exceed human-level performance in selected (narrowly- or ill-defined) tasks, with the aim to fully automate and replace humans. This goal has nonetheless repeatedly proven difficult to attain due to brittle predictions whose subpar fairness, interpretability and robustness as well as ambiguous accountability are concerning, especially given their potential harm. By considering the broader organisational and societal context in which data-driven systems are operationalised, we should not only strive to automate and replace (when appropriate and desirable) but also to augment and support human reasoning and decision-making to help people flourish at work, e.g., through human–machine collaboration that preserves people's agency and maintains the attribution of responsibility with them. Such a perspective promises to offer an antidote to widely reported apprehension of artificial intelligence and expedite its adoption in safety critical domains.

Seminar Topic

To address these challenges, our interdisciplinary seminar gathers a broad range of stakeholders – including clinicians, academics and researchers from industry – whose diverse expertise can contribute to charting a novel research agenda for effective and responsible adoption of artificial intelligence in medicine given the complex sociotechnical landscape outlined above. Our goal is to identify best ways of operationalising medical data-driven systems as to ensure their alignment with the needs and expectations of various stakeholders in healthcare as well as seamless integration into real-life clinical workflows, taking a human-centred perspective. Exploring these aspects of artificial intelligence is especially important given that achieving state-of-the-art performance on benchmark tasks often does not directly translate into clinical efficacy and acceptability. To support this objective, we additionally intend to scrutinise relevant evaluation procedures, medical device certification processes, practicality of clinical trials involving data-driven algorithms and clinical approvals thereof in view of compliance with various laws, rules and regulations as well as societal norms and ethical standards. Throughout the seminar we envisage identifying challenges that can be addressed with current technologies, distilling areas that require further work, and emphasising promising research directions. Finally, the event aims to galvanise an interdisciplinary community dedicated to advancing the meeting's agenda after its conclusion.

Seminar Outcomes

The seminar focused on the challenges of translating medical artificial intelligence (AI) models from research settings to real-world clinical applications. It brought together academic and

industry researchers, start-up representatives as well as practising clinicians to foster a multidisciplinary exchange of ideas. One of the key highlights of the seminar was an invited keynote by Rich Caruana from Microsoft Research. His presentation on ante-hoc interpretable models emphasised the importance of intelligibility in machine learning for healthcare. This talk sparked significant discussions among the participants and served as a catalyst for many of the conversations that followed.

Throughout the seminar, the participants engaged in a variety of discussions and presentations. Clinicians were invited to share their experiences with data-driven decision support systems, focusing on both success stories and ongoing challenges; they were also encouraged to describe their hopes and vision for the future of such tools. These clinical pitches played a central role in shaping the seminar's core themes, which included research, translation, testing, deployment, monitoring, updating and maintenance of AI systems in healthcare. Additionally, researchers delivered short presentations on their work, providing insights into the state of the art as well as open research problems in clinical AI systems. A dedicated session for start-ups offered valuable insights into the process of transforming research findings into real-life clinical tools. Among others, entrepreneurs shared their experiences with commercialisation and the regulatory hurdles they encountered. Many discussions revolved around the practical aspects of deploying AI in healthcare settings and the lessons learnt from these experiences.

The seminar also facilitated group work; two dedicated working groups were formed. The first group focused on frameworks for evaluation and (post-deployment) monitoring of clinical AI. The second group explored important criteria to consider when selecting clinical problems for which to develop AI tools; it additionally investigated human factors of medical AI systems and key approaches to improve the interaction between AI and doctors.

Overall, the seminar identified pressing challenges and opportunities in clinical AI research and deployment. Clinicians gained a deeper understanding of AI's capabilities and limitations, while researchers benefited from the exchange of strategies for overcoming integration and adoption barriers. The discussions and findings from the seminar are expected to facilitate smoother transitions from research to clinical AI prototypes, allowing such tools to be tested and deployed in hospitals. By fostering interdisciplinary collaboration, the seminar laid the groundwork for future innovations in AI-driven clinical decision support systems. The insights shared and connections formed during the event will contribute to ongoing advancements in the field and help bridge the gap between AI research and practical healthcare applications.

2 Table of Contents

3 Overview of Talks

3.1 Bridging the Gap Between Clinical Data and AI: Lessons From Real-World EMR Studies

Brett Beaulieu-Jones (University of Chicago, US)

Healthcare data, particularly electronic medical records (EMRs), present significant challenges due to their complexity, inconsistency and inherent biases. This presentation explores the implications of these issues for clinical artificial intelligence (AI) and phenotyping models, emphasising the role of clinician-initiated (CI) versus non-clinician-initiated (NCI) data in predictive modelling. Using real-world case studies, we examine how AI models interpret EMR data, the risks of confounding feedback loops in clinical decision support and the divergence between models trained on CI and NCI data. We highlight findings from large-scale EMR studies on patient risk stratification, model performance limitations and the impact of institutional effects. Additionally, we discuss the dangers of label leakage, reproducibility challenges in published predictive models and the unintended consequences of AI-based clinical alerts. The talk underscores the need for rigorous evaluation of AI models deployed in clinical settings to ensure they enhance, rather than hinder, medical decision-making.

3.2 Implementing Clinical Workflows in the Clinic

Michael Brudno (University of Toronto, CA)

In this presentation I will look at the challenges of implementing machine learning (ML) in a hospital setting, concentrating specifically on integrating ML into clinical workflows in a safe and effective manner. I will utilise two examples from my research: the deployment of Machine Learning Medical Directives (MLMD) for making low-risk decision in paediatric Emergency Rooms and scheduling of craniosynostosis and plagiocephaly patients for Plastic Surgery consultations based on their likely risk and urgency.

To develop MLMD we used data from the EHR system from the Hospital for Sick Children, a tertiary care hospital in the city of Toronto, Canada to train multiple ML models to predict the need for urinary dipstick testing, ECGs, abdominal ultrasounds, testicular ultrasounds, bilirubin testing and forearm X-rays using data available at triage. There was a total of 42,238 patients (54.7% boys) included in model development; mean (SD) age of the children was 5.4 (4.8) years. Models obtained high area under the receiver operator curve (0.89–0.99) and positive predictive values (0.77–0.94) across each of the use cases. The proposed implementation of MLMDs would streamline care for 22.3% of all patient visits and make test results available earlier by 165 minutes (weighted mean) per affected patient. Model explainability for each MLMD demonstrated clinically relevant features having the most influence on model predictions. In the presentation we emphasised the safety of deploying these ML models and the importance of considering clinical workflows (staff availability, importance of explaining the AI models to patients, etc.) in deployment.

In the second example we consider the scheduling of appointments of craniosynostosis and plagiocephaly patients in a plastic surgery department. Craniosynostosis is a birth defect that results in a misshapen skull due to premature bone fusion as a newborn's skull is formed. In some cases, skulls with this defect do not have adequate space for the newborn's brain to grow, which increases the chance of visual and mental development impairments; almost all cases of craniosynostosis also result in head shape abnormalities that may lead to bullying and impact individual self-perception. Craniosynostosis can be corrected by relatively non-invasive surgery before 3 months; after this age, however, patients require more complex surgery with higher morbidity. Craniosynostosis is typically diagnosed by a physical examination by a specialist, such as a paediatric plastic surgeon. Paediatricians who are not trained at identifying craniosynostosis often confuse it for plagiocephaly, a related but mostly benign condition, and typically refer patients with either condition to plastic surgery for a definitive diagnosis. However, the delay associated with the referral process can require the more complex surgical approach. We have recently demonstrated that 3D head shape reconstruction using a standalone ToF camera (3DMD system) can aid in the identification of craniosynostosis with high accuracy and allow prioritisation of referred patients. Again, deploying this tool into clinical care requires careful consideration of existing clinical workflows. While reducing the overall burden for some families, it would require patients to make multiple visits (one to have a 3D photo taken, one to see the surgeon for a comprehensive evaluation). This would potentially lead to some inequity, as patients further from the hospital would require more resources to benefit from the AI.

In discussing both cases I will emphasise the important "fall-back" mechanisms, where if a patient is not flagged by the ML system, they will still undergo standard-of-care treatment, and also consider how the presence of automation may impact clinicians who become "accustomed" to having the support, and may fail to act appropriately if the technology malfunctions.

3.3 Friends Don't Let Friends Deploy Black Box Models: The Importance of Intelligibility in Machine Learning for Healthcare

Rich Caruana (Microsoft – Redmond, US)

The conventional wisdom in machine learning has been that to achieve high accuracy you must use opaque black-box models such as deep neural nets, boosted trees or random forests, and that if you want models to be interpretable and able to explain their predictions, you have to accept a loss in accuracy. This trade-off is no longer true when working with tabular data – in the last 10 years glass-box learning methods have been developed that are just as accurate as black-box learning methods but which are fully interpretable and can explain their predictions. Applying these glass-box learning methods to healthcare data has uncovered many problems inherent in clinical data that would make models trained on the data risky to use on patients. These problems include selection bias, race and gender bias, treatment effects, other forms of statistical confounding and problems with popular methods of dealing with missing data and data coding.

None of these are new problems. What is new is how widespread these problems are, how unexpected some of them are even in high-quality well-curated data and the difficulty of correcting these problems using traditional methods. The new high-accuracy glass-box

learning methods have shown that these problems exist in every dataset. Moreover, these problems make all black-box models trained on medical data suspect because one is unable to anticipate all of the problems in advance and it is difficult to fully understand after the fact what has been learnt by complex black-box models. Glass-box learning methods not only make it easier to detect these problems, but also provide tools for correcting many of these problems by allowing clinicians to use their expertise to directly correct/edit the models when they have learnt patterns that would put patients at risk.

In the talk we examined a half dozen case studies using real medical data that show the kinds of problems that are common in medical datasets, and how we would use glass-box learning to detect and then correct these problems. The case studies serve as a wake-up call to anyone using machine learning and artificial intelligence in healthcare that if they are training and/or using models that they cannot fully understand (i.e., black-box models), then they are almost certainly putting patients at higher risk if model predictions are acted upon. In addition to providing models that are fully interpretable, some of the new glass-box learning methods not only provide methods to correct models, but also can explain their reasoning and help protect privacy. Now that glass-box learning methods are so powerful, it would be wrong to intentionally use black-box models in critical domains such as healthcare if glass-box models yielded comparable accuracy.

The talk was not about the technical details of any one glass-box learning method. Instead, it was a collection of case studies that show the dangers of using black-box models in healthcare and how glass-box methods can be used to mitigate these risks. Once the problems hidden in each dataset are uncovered, there may be multiple methods available to tackle the problems, but the key challenge is to detect the problems in the first place so that they can be corrected prior to deploying the model.

3.4 Validation in Biomedical Imaging AI: Are We Ready for Clinical Translation?

Evangelia Christodoulou (DKFZ – Heidelberg, DE)

Reliable validation of machine learning (ML) algorithms remains a critical challenge, particularly in biomedical image analysis, where chosen performance metrics often fail to reflect domain interests. To address this, we introduce Metrics Reloaded, a comprehensive framework guiding researchers in selecting problem-aware validation metrics. Developed by an international consortium, it employs a structured problem fingerprint to capture key aspects influencing metric selection. Additionally, we highlight a crucial limitation in current performance reporting: the widespread neglect of performance variability. Analysing 221 MICCAI 2023 segmentation papers, we find that over 50% do not assess variability, and only 0.5% report confidence intervals (CIs). To bridge this gap, we propose an approximation method that reconstructs CIs using unreported standard deviation values, revealing that reported performance differences often lack statistical significance. Together, these contributions aim to enhance ML validation practices, ensuring more reliable and clinically relevant algorithm evaluation.

3.5 Towards Deployment: Considerations Beyond Technical Performance

Jeff Clark (IngeniumAI – Bath, GB)

When developing a decision support system, it is tempting to focus most of your energy on the core technology: the technical innovation, which we believe will have a positive impact on the healthcare system once deployed. In this talk I touch upon many of the other required facets, which must be pursued in parallel, as you move from a research project towards deployment. This includes technical considerations concerned with safe deployment such as prospective performance and drift, but also many factors beyond the performance of the core technology, including but not limited to: initiating a quality management system, regulatory evidence and documentation, route to market strategy, human factors and healthcare economics. None of these other factors can be ignored, and will be pivotal to the success of deploying your innovation.

3.6 AI at the Bedside. There Must Be a Culture Change

James Fackler (Johns Hopkins University – Baltimore, US)

Focused on paediatric critical care, I believe there are no current uses of machine learning (ML) or artificial intelligence (AI) in general that have reached the bedside. However, I remain optimistic that AI will have a profound impact on patient care in the next five years (or ten at the longest).

To leverage AI at the bedside will require a substantial medical culture change. Because knowledge will be "ubiquitous", the traditional hierarchy where the doctor (or in academic medicine, the attending physician) is the final arbiter of truth and the sole source of a care plan, will be upended. Patients will have access to the same knowledge as do the doctors. The role of the senior clinician will become one who understands what AI "knows" and more importantly what AI does not (or cannot) know. Individuals on the care team (e.g., nurses, junior physicians, pharmacists) will develop the same relationship with AI and knowledge but will do so within the "niche" expertise.

3.7 A Few Lessons Learnt From Trying To Work With Healthcare and Related Data

Thomas Gärtner (Technische Universität Wien, AT)

In my talk, I reported on a variety of experiences from (so far) mostly unsuccessful attempts of applying machine learning algorithms to healthcare and related data. My first experience was with time series of oxygen levels taken during brain surgeries and was available for

a very small number of patients only. My next experience was on images of eyes with implanted lenses for cataract patients and could be solved sufficiently well without the use of sophisticated machine learning algorithms. My most recent experience is with clinical studies and involves long discussions about NDAs and IPRs with legal departments.

3.8 AI in Healthcare: Key Human–Computer Interaction Challenges

Maia Jacobs (Northwestern University – Evanston, US)

The use of artificial intelligence (AI) for improving medical decision-making has garnered great excitement in recent years. Yet, despite growing enthusiasm and increased research, real-world clinical impact has been slow. Often, abandonment of these tools in clinical settings is not related to algorithmic performance, but rather due to inattention towards the technologies' design and implementation. To understand and address these challenges, I will share two of my lab's research projects, which use user-centred and participatory design methods to incorporate both providers' and patients' perspectives into clinical decision support systems.

3.9 From Code to Clinic – From Bits to Bedside

Michael Kamp (Universitätsmedizin Essen, DE)

The integration of artificial intelligence (AI) into clinical practice requires not only technological advancements but also rigorous methods to ensure reliability, privacy and interpretability. At the University Hospital Essen (UK Essen), one of the world's leading smart hospitals, the Institute for AI in Medicine (IKIM) develops and deploys AI systems within a large-scale data infrastructure based on Europe's largest FHIR server. This enables advanced machine learning applications while maintaining strict data governance.

This talk will present research from the Trustworthy Machine Learning group, focusing on three core challenges in medical AI: privacy-preserving federated learning, where we move beyond standard model aggregation techniques to improve learning from distributed clinical data; statistical performance guarantees, leveraging theoretical insights from loss surface analysis to better understand generalisation in deep learning; and (federated) causal discovery, which aims to disentangle causal relationships in medical datasets to improve model interpretability and robustness.

By combining these approaches, we work toward AI models that are not only predictive but also scientifically grounded and reliable in clinical decision-making. The talk will discuss recent advancements in federated learning, causal inference and generalisation theory, along with their implications for AI applications in healthcare.

3.10 Learning From Machine Learning – How To Deduce a Mechanism-Based Pharmacometrics Model for Serum Creatinine in Preterm Neonates From Neural Ordinary Differential Equations

Gilbert Koch (Universitäts-Kinderspital beider Basel, CH)

Introduction

Machine learning (ML) is an emerging field in pharmacometrics (PMX) [1], providing methods for a variety of PMX tasks, including data preparation [2], data analysis and data modelling. One ML approach gaining special attention in PMX are neural ordinary differential equations (NODEs) [3, 4, 5, 6]. Although an NODE is basically an ordinary differential equation (ODE), the difference is that the right-hand side of the ODE is not described with mechanism-based functions, as it is typically done in PMX, but it consists of neural networks (NNs). Consequently, these NNs learn the dynamics observed in the training data. However, there are some major criticisms regarding NODEs, including that (i) they are "black box" models, (ii) they have poor extrapolation capabilities, e.g., for unseen dose ranges, due to their structure, and (iii) they do not include prior clinical knowledge. In this work, a reverse modelling approach is presented that leverages the learnt knowledge by a NODE to deduce a mechanism-based model allowing to additionally include clinical knowledge. This enables to overcome the criticism of NODEs mentioned above and to make them a more viable approach in the field of PMX.

Methods

As endurance test, a dataset consisting of serum creatinine concentration measurements ($n = 4{,}026$) from extremely low birth weight neonates ($n = 217$) with marked renal maturation processes was applied [7]. The low-dimensional NODE approach was utilised [6] where the right-hand side of the NODE consists of two types of NNs specifically tailored to PMX. The first NN takes the state as input, reflecting the autonomous behaviour of the dynamics. The second NN takes explicit time as input, reflecting behaviour of the dynamics that change over time, e.g., maturation processes. First, the serum creatinine measurements were fitted with the low-dimensional NODE in the non-linear mixed-effects context in Monolix and a covariate analysis was performed. Second, the learnt dynamics of the NNs were visualised in derivative versus state or time plots [6]. Based on visual inspection of these plots, PMX functions were selected that described the shape of the trajectories in these plots. Third, these PMX functions were combined to deduce a mechanism-based model that is capable to characterise the dynamics of serum creatinine concentrations. Fourth, this deduced mechanism-based model was further refined with clinical knowledge about the influence of body weight on the volume of distribution. As last step, this deduced final mechanism-based model was fitted to the data, a covariate analysis was conducted with the previously gained information from the NODE-covariate analysis and simulations were performed.

Results

The developed low-dimensional NODE was capable of learning complex dynamics of serum creatinine in preterm neonates with good measures of precision and bias (mean squared error MSE = 0.023 and relative mean prediction error RMPE = 1.471). In comparison to the

previously published model [7], the NODE model provided similar data fitting and simulated similar GA-dependent reference values. Further it was able to identify the most important covariates found in the previously published model. Based on the visualised trajectories in the derivative versus state or time plots, a linear function for the NN characterising the state and an Emax function for the NN describing time were chosen. Remarkably, the deduced mechanism-based model had a similar structure as the previously published serum creatinine model [7]. In addition, clinical knowledge was included, i.e., volume of distribution for serum creatinine was assumed to be 7 dL/kg, resulting in the final mechanism-based model with similar measures of precision and bias as the NODE model (MSE = 0.025, RMPE = -2.17). It should be noted that NODE-based ML approach dramatically reduced time effort associated with the development of a mechanism-based model describing serum creatinine dynamics in neonates.

Conclusion

A mechanism-based model was successfully deduced from the dynamics learnt by the NODE. Structure of the deduced mechanism-based model was in accordance with a previously published, conventionally developed model for serum creatinine concentration in preterm neonates. Hence, we demonstrated the potential that initially learning the dynamics by an NODE is expected to accelerate development of mechanism-based models, particularly in paediatrics.

References

1 Alexander Janssen, Frank C. Bennis, and Ron A. A. Mathôt. *Adoption of machine learning in pharmacometrics: An overview of recent implementations and their considerations*. Pharmaceutics, 14(9), p.1814, MDPI, 2022

2 Dominic Stefan Bräm, Uri Nahum, Andrew Atkinson, Gilbert Koch, and Marc Pfister. *Evaluation of machine learning methods for covariate data imputation in pharmacometrics*. CPT: Pharmacometrics & Systems Pharmacology, 11(12), p.1638–1648, Wiley Online Library, 2022

3 Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. *Neural ordinary differential equations*. Advances in Neural Information Processing Systems, 31, 2018

4 James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. *Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens*. iScience, 24(7), Elsevier, 2021

5 Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. *Universal differential equations for scientific machine learning*. arXiv preprint arXiv:2001.04385, 2020

6 Dominic Stefan Bräm, Uri Nahum, Johannes Schropp, Marc Pfister, and Gilbert Koch. *Low-dimensional neural ODEs and their application in pharmacokinetics*. Journal of Pharmacokinetics and Pharmacodynamics, 51(2), p.123–140, Springer, 2024

7 Tamara van Donge, Karel Allegaert, Verena Gotta, Anne Smits, Elena Levtchenko, Djalila Mekahli, John van den Anker, and Marc Pfister. *Characterizing dynamics of serum creatinine and creatinine clearance in extremely low birth weight neonates during the first 6 weeks of life*. Pediatric Nephrology, 36, p.649–659, Springer, 2021

3.11 AI for Chronic Care Management

Yamuna Krishnamurthy (Phamily – New York, US)

In this talk I presented how we, at Phamily, are empowering chronic care management with artificial intelligence (AI). Phamily is a healthcare start-up with a vision to provide value-based care to chronic care patients. Our goal is to provide a system that brings physicians, nurses, care managers and patients together for continued conversations about the care that the patients need and how they can be addressed by the medical staff in between office visits. AI is the much needed assistant to the care managers that can help them quickly do chart reviews, draw up care plans and engage the patients. It can also assess short- and long-term patient risks for early detection and timely intervention that can save lives and prevent exorbitant costs for all involved.

3.12 Rethinking Medical AI: Evaluation, Representation and Transferability

Christoph Lippert (Hasso-Plattner-Institut, Universität Potsdam, DE)

Medical artificial intelligence (AI) systems promise to revolutionise clinical practice, but questions about their real-world effectiveness and interoperability remain largely unanswered. In this talk, I addressed two fundamental questions: *Do we need to evaluate medical AI?* and *Do we need ontologies?*

In the first part, I discussed insights from our study on commercial AI systems for tuberculosis detection [1], where we found that key information – such as training population details – is often lacking or opaque. This undermined model applicability, especially in global health settings, and required us to conduct extensive pilot testing to adapt a commercial algorithm for use in a South African community-based screening initiative.

In the second part, I turned to the role of ontologies in medical AI. Based on our recent work [2], I argued that representations learnt by large language models (LLMs) offer a superior and more scalable alternative to traditional medical ontologies. Our GRASP model embeds medical codes into a unified semantic space using LLMs, enabling cross-system and cross-country transferability of EHR-based prediction models – even without harmonised data models.

Taken together, the talk advocates for greater transparency in evaluation and a shift from static ontologies to dynamic, data-driven language representations for advancing trustworthy and generalisable medical AI.

References

1 Jana Fehr, Stefan Konigorski, Stephen Olivier, Resign Gunda, Ashmika Surujdeen, Dickman Gareta, Theresa Smit, Kathy Baisley, Sashen Moodley, Yumna Moosa, Willem Hanekom, Olivier Koole, Thumbi Ndung'u, Deenan Pillay, Alison D. Grant, Mark J. Siedner, Christoph Lippert, Emily B. Wong, and Vukuzazi Team. *Computer-aided interpretation of chest radiography reveals the spectrum of tuberculosis in rural South Africa*. npj Digital Medicine, 4(1), p.106, Nature Publishing Group UK London, 2021

2 Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, FinnGen, Christoph Lippert, and Andrea Ganna. *Large language models improve transferability of electronic health record-based predictions across countries and coding systems*. medRxiv, 2025–02, Cold Spring Harbor Laboratory Press, 2025

3.13 All Models Are Wrong and Yours Are Useless

Florian Markowetz (University of Cambridge, GB)

Most published clinical prediction models are never used in clinical practice and there is a huge gap between academic research and clinical implementation. In this talk I propose ways for academic researchers to be proactive partners in improving clinical practice and to design models in ways that ultimately benefit patients.

3.14 Predictive Analytics Monitoring at the Bedside

Randall Moorman (University of Virginia – Charlottesville, US)

Predictive analytics monitoring is a new field of research and development that is ready for clinical implementation. The precepts are that there are detectable signatures of illness in continuous monitoring data and that detection of these signatures can lead to early detection, early treatment and improved outcomes.

I describe a successful example. Sepsis in premature infants is a common and pernicious problem but, if the culprit infection is diagnosed early, antibiotic treatment averts severe morbidity and mortality. We found more than 20 years ago that there was a robust signature of illness, reduced variability and transient decelerations of heart rate, that appeared hours before clinical presentation. A very large randomised trial showed that infants with a display of a risk index based on these abnormal heart rate characteristics had improved all-cause survival. While the signature was detected by visual inspection of many heart rate records by clinicians, the same signature was detected by machine learning and deep learning methods.

There are challenges to this new field. Before modelling begins, data sets may not have well-annotated target events and the data may reflect the clinicians and not the patients. When modelling, deep learning may be no better than machine learning, bias in the data may lead to bias in the model and the model output may not be explainable to clinicians. After deployment, effective implementation is difficult and the model will not work the same if the data shift.

3.15 Interpretability?

Rajesh Ranganath (NYU Courant Institute of Mathematical Science, US)

Joint work of Aahlad Puli, Nhi Nguyen, Rajesh Ranganath

Feature attributions attempt to highlight what inputs drive predictive power. Good attributions or explanations are thus those that produce inputs that retain this predictive power; accordingly, evaluations of explanations score their quality of prediction. However, evaluations produce scores better than what appears possible from the values in the explanation for a class of explanations called encoding explanations. Probing for encoding remains a challenge because there is no general characterisation of what gives the extra predictive power. We develop a definition of encoding that identifies this extra predictive power via conditional dependence and show that the definition fits existing examples of encoding. This definition implies, in contrast to encoding explanations, that non-encoding explanations contain all the informative inputs used to produce the explanation, giving them a "what you see is what you get" property, which makes them transparent and simple to use. Next, we prove that existing scores (ROAR, FRESH, EVAL-X) do not rank non-encoding explanations above encoding ones, and develop STRIPE-X, which ranks them correctly. After empirically demonstrating the theoretical insights, we use STRIPE-X to show that despite prompting a large language model (LLM) to produce non-encoding explanations for a sentiment analysis task, the LLM-generated explanations encode.

3.16 AI in Paediatric Surgery and Paediatric Urology

Patricia Reis Wolfertstetter (KH Barmh. Brüder Klinik St. Hedwig – Regensburg, DE)

Artificial intelligence and machine learning models are promising tools for the further development of paediatric surgery and paediatric urology. They can be used for optimising treatment and patient stratification preoperatively, during operation and postoperatively. Up to now, our work focused on paediatric appendicitis. First, predictive models were developed and validated on a dataset acquired from 430 children and adolescents aged 0–18 years, based on a range of information encompassing history, clinical examination, laboratory parameters and abdominal ultrasonography. Logistic regression, random forests and gradient boosting machines were used for predicting the three target variables: diagnosis, treatment and severity. Furthermore, we presented interpretable machine learning models for predicting the diagnosis, management and severity of suspected appendicitis using ultrasound images. Our approach utilised concept bottleneck models (CBM) that facilitate interpretation and interaction with high-level concepts understandable to clinicians. We extended CBMs to prediction problems with multiple views and incomplete concept sets. Our models were trained on a dataset comprising 579 paediatric patients with 1,709 ultrasound images accompanied by clinical and laboratory data. The developed models are deployed as an open access easy-to-use online tool (for tabular data and ultrasound images).

3.17 My Priorities for AI in Health: Proper Evaluation and Prediction Under Intervention

Wouter van Amsterdam (University Medical Center Utrecht, NL)

Artificial intelligence systems in healthcare must ultimately support safe and effective decision-making. In this talk, I argue that evaluation should extend beyond predictive performance on held-out data to the real-world setting – treating deployment itself as an intervention. I highlight how misaligned evaluation metrics can lead to harmful self-fulfilling prophecies, especially in treatment decision support.

Next, I argue researchers should build models for "prediction under intervention" (sometimes referred to as counterfactual prediction): estimating what would happen under different treatment options rather than expected outcomes under historical regimes. I draw on methods from causal inference and off-policy evaluation, and reflect on how emerging regulatory frameworks (EU and FDA) and available randomised control trial data can support a more rigorous approach.

3.18 Scaling up Clinical ML: Modalities, External Validation, Health Systems

Robin Van de Water (Hasso-Plattner-Institut, Universität Potsdam, DE)

Scaling medical machine learning (ML) requires integrating multiple modalities, improving model validation practices and establishing robust infrastructure to handle massive amounts of clinical data.

Scaling Up to Different Modalities

In visceral surgery, postoperative complications often arise in nursing wards, where real-time monitoring is limited. While ML-powered predictive systems show promise in the intensive care unit (ICU), their effectiveness diminishes outside of it due to data shortages, leaving patients at risk. To address this, we propose an integrated approach that combines patient data from preoperative, intraoperative, ICU and nursing ward stages, while introducing high-resolution continuous vital sign monitoring in a hybrid nursing environment [2]. This system enhances early detection of complications like surgical site infections and bile leakage, demonstrating the importance of high-quality wearable data. Our findings suggest that hybrid monitoring can significantly improve ML-based early warning models in clinical settings, enhancing patient outcomes.

Scaling Up External Validation

One of the biggest challenges in scaling clinical ML is ensuring reproducibility and transparency across datasets. With ICU models, it is difficult to verify claims of superior performance due to lack of access to datasets as well as unclear cohort definitions and preprocessing steps. To address this, we introduce YAIB, a modular framework designed to support reproducible

clinical ML experiments with multiple open-access ICU datasets [3]. YAIB provides an end-to-end solution for model evaluation, including predefined tasks like mortality and sepsis, and highlights the critical role of dataset selection, cohort definition and preprocessing in model performance. By offering a unified benchmarking tool, YAIB paves the way for more transparent, comparable ML research in clinical settings.

Scaling Up to Entire EHR Systems

For ML to be deployed at scale across entire health systems, we need an efficient infrastructure to process both retrospective and prospective clinical data. To meet this need, we developed the Medical Event Data Standard (MEDS) [1], a flexible low-level ML standard that integrates seamlessly with existing electronic health record processing and modelling tools. MEDS accelerates the training of predictive models on several clinical tasks. As a proof of concept, we built an ETL pipeline to enable model development using the recently released NWICU dataset. Additionally, we are working to convert data from the Mount Sinai AIR MS PHI OMOP database into MEDS, making it possible to train models for various clinical endpoints and develop foundation models. With ongoing efforts to verify data quality and define cohorts, we aim to enhance model robustness and further advance the scalability of ML in healthcare.

References

1 Bert Arnrich, Edward Choi, Jason Alan Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. *Medical event data standard (MEDS): Facilitating machine learning for health*. ICLR 2024 Workshop on Learning from Time Series For Health, 2024

2 Robin van de Water, Axel Winter, Max M. Maurer, Felix August Treykorn, Daniela Zuluaga, Bjarne Pfitzner, Igor M. Sauer, and Bert Arnrich. *Combining hospital-grade clinical data and wearable vital sign monitoring to predict surgical complications*. ICLR 2024 Workshop on Learning from Time Series For Health, 2024

3 Robin van de Water, Hendrik Nils Aurel Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. *Yet another ICU benchmark: A flexible multi-center framework for clinical ML*. Proceedings of the Twelfth International Conference on Learning Representations, 2023

3.19 AI in Babies and Beyond, Boom or Boomerang?

Sven Wellmann (Universität Regensburg, DE)

Birth is one of the most critical moments in a person's life. Birth marks the transition from life in the womb (pregnancy) to life outside the womb. The infant's survival and growth depend fundamentally on basic support for months and years. Many disorders affecting the nervous system lifelong originate in early life and in particular in perinatal complications such as neonatal encephalopathy, preterm birth, neonatal sepsis or jaundice.

We will learn how medical examinations of newborn babies are routinely performed immediately after birth and during the first weeks out of the womb, how vital signs indicate healthy body functions and how subtle clinical signs can point towards serious problems that require more complex examination and possibly subsequent treatment. Building on

this, we will discuss areas in the care of pregnant women and newborns where improved diagnostics through the use of computer algorithms could contribute to reducing morbidity and mortality.

The introduction of prediction algorithms and decision support tools based on methods of so-called artificial intelligence (AI) has started in neonatology and paediatrics. We will discuss first use cases and possible benefits in reducing healthcare gaps. However, we will also shed a light on potential risks that may harm the baby's well-being despite the current boom in AI.

4 Working Groups

4.1 AI Monitoring in Clinical Practice

Brett Beaulieu-Jones (University of Chicago, US), Evangelia Christodoulou (DKFZ – Heidelberg, DE), Thomas Gärtner (Technische Universität Wien, AT), Michael Kamp (Universitätsmedizin Essen, DE), Gilbert Koch (Universitäts-Kinderspital beider Basel, CH), Yamuna Krishnamurthy (Phamily – New York, US), Fabian Laumer (Scanvio Medical AG, CH), Christoph Lippert (Hasso-Plattner-Institut, Universität Potsdam, DE), Florian Markowetz (University of Cambridge, GB), Randall Moorman (University of Virginia – Charlottesville, US), Rajesh Ranganath (NYU Courant Institute of Mathematical Science, US), Raul Santos-Rodriguez (University of Bristol, GB), Wouter van Amsterdam (University Medical Center Utrecht, NL), Robin Van de Water (Hasso-Plattner-Institut, Universität Potsdam, DE), and Julia E. Vogt (ETH Zürich, CH)

Our working group has identified a fundamental challenge in post-deployment monitoring of clinical artificial intelligence (AI) systems: a misalignment between incentives and resources that undermines effective oversight. Those with the greatest interest in ensuring AI safety and efficacy – clinicians, researchers and patients – often lack the necessary funding, technical infrastructure and institutional support. Meanwhile, AI vendors and large healthcare systems, which have these resources, frequently lack strong incentives to engage in long-term monitoring.

To address this issue and align stakeholder interests, introducing regulatory mandates, standardised monitoring metrics and financial incentives may be necessary. Creating clear reporting requirements, real-world performance evaluations and publicly accessible monitoring databases appears particularly promising for enhancing transparency and trust in clinical AI tools. Achieving these objectives will likely require tight collaboration between regulators, developers and healthcare providers, with the goal of establishing best practices and ensuring continuous oversight. Without such measures, AI-driven healthcare solutions risk inconsistent safety and effectiveness, ultimately limiting their long-term benefit to patient care.

4.2 Human Factors in Clinical AI Design and Deployment

Michael Brudno (University of Toronto, CA), Jeff Clark (IngeniumAI – Bath, GB), James Fackler (Johns Hopkins University – Baltimore, US), Maia Jacobs (Northwestern University – Evanston, US), Patricia Reis Wolfertstetter (KH Barmh. Brüder Klinik St. Hedwig – Regensburg, DE), Kacper Sokol (ETH Zürich, CH), and Sven Wellmann (Universität Regensburg, DE)

Our working group reviewed key criteria for selecting clinical problems where artificial intelligence (AI) can have the most meaningful impact; we also focused on human factors that are fundamental to ensuring safe and effective deployment of AI in clinical practice. One important point is the role of biological plausibility. Should AI operation always align with known and accepted medical knowledge, or can models be trusted even if these mechanisms are unclear? Additionally, clinician trust tends to depend on both accuracy and explainability of AI, but how much weight should be given to each remains an open question.

Another key consideration is how to integrate AI into clinical workflows. Can AI systems be adapted to existing medical workflows, or should they be designed to drive (beneficial) workflow changes over time? Crucially, AI could play a role in operational improvements – such as staffing predictions and workflow optimisation – while balancing feasibility and impact. Also, the characteristics of the clinical challenge for which an AI solution is envisaged need to be considered. For example, should AI development focus on supporting (and possibly automating) routine decisions, allowing clinicians to devote their attention to more complex cases? Moreover, we explored opportunities for AI chatbots in collection of patient history, provision of feedback to clinicians and general decision support; nonetheless, questions remain about how to define their limits and handle sensitive topics.

From the human factors perspective, ensuring graceful AI failure modes and designing intuitive handover protocols are of paramount importance so that operators can easily identify such cases and handle them appropriately. A related issue is the need for AI to recognise when it encounters unfamiliar cases and transition control back to human oversight without disrupting provision of care. Additionally, where and how AI-generated alerts should appear in a clinician's workflow remains an open question. Their role is also unclear: should they be advisory, mandatory or something in-between?

Past failures of deployed clinical AI systems highlight the risks of over-reliance on these tools, especially without clear understanding of their limitations. Moreover, we need to consider professional and cultural barriers. For example, how can we ensure that AI benefits all types of healthcare professionals, from nurses to senior physicians? In this context, another important open question is how AI can facilitate teamwork, particularly in patient handoff between clinicians; should AI simply provide information, or should it actively suggest next steps?

Participants

- Brett Beaulieu-Jones
University of Chicago, US
- Michael Brudno
University of Toronto, CA
- Evangelia Christodoulou
DKFZ – Heidelberg, DE
- Jeff Clark
IngeniumAI – Bath, GB
- James Fackler
Johns Hopkins University – Baltimore, US
- Thomas Gärtner
Technische Universität Wien, AT
- Maia Jacobs
Northwestern University – Evanston, US
- Michael Kamp
Universitätsmedizin Essen, DE
- Gilbert Koch
Universitäts-Kinderspital beider Basel, CH
- Yamuna Krishnamurthy
Phamily – New York, US
- Fabian Laumer
Scanvio Medical AG, CH
- Christoph Lippert
Hasso-Plattner-Institut, Universität Potsdam, DE
- Florian Markowetz
University of Cambridge, GB
- Randall Moorman
University of Virginia – Charlottesville, US
- Rajesh Ranganath
NYU Courant Institute of Mathematical Science, US
- Patricia Reis Wolfertstetter
KH Barmh. Brüder Klinik St. Hedwig – Regensburg, DE
- Raul Santos-Rodriguez
University of Bristol, GB
- Kacper Sokol
ETH Zürich, CH
- Wouter van Amsterdam
University Medical Center Utrecht, NL
- Robin Van de Water
Hasso-Plattner-Institut, Universität Potsdam, DE
- Julia E. Vogt
ETH Zürich, CH
- Sven Wellmann
Universität Regensburg, DE