DAGSTUHL
REPORTS

Volume 15, Issue 7, July 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
April, 2026

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

Digital Object Identifier: 10.4230/DagRep.15.7.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 25281

From Sparse Interpolation to Signal Processing: New Synergies

Annie Cuyt[*1], Dirk de Villiers[*2], Wen-shin Lee[*3], Ana C. Matos[*4], Gerlind Plonka-Hoch[*5], and Ramonika Sengupta[†6]

1 **University of Antwerp, BE.** annie.cuyt@uantwerpen.be
2 **University of Stellenbosch, ZA.** ddv@sun.ac.za
3 **University of Stirling, GB.** wen-shin.lee@stir.ac.uk
4 **University of Lille, FR.** ana.matos@univ-lille1.fr
5 **University of Göttingen, DE.** plonka@math.uni-goettingen.de
6 **Eindhoven University of Technology, NL.** r.sengupta@tue.nl

Abstract

In a data-rich digital world, finding sparse, efficient representations – especially for multi-exponential models – has become critical, particularly when measurements are costly or noisy. These models, which involve complex or real exponents, underpin key processes in signal processing, relaxation dynamics, chemical reactions, heat transfer, and fluid dynamics, with widespread real-world impact. The challenge lies at the intersection of several computational disciplines: structured matrices, rational approximation, sparse interpolation, quadrature, tensor decompositions, and subdivision methods – each offering potential pathways to more robust and efficient algorithms. Multi-exponential analysis is foundational across engineering and industry, enabling advances in DOA estimation, remote sensing, MRI, superresolution, seismology, radio astronomy, and telecommunications – areas vital to energy, health, transportation, and space research. This Dagstuhl Seminar "From Sparse Interpolation to Signal Processing: New Synergies" (25281) brought together experts from computational harmonic analysis, numerical linear algebra, computer algebra, signal processing, approximation theory, and engineering applications to foster cross-disciplinary collaboration and accelerate innovation in this dynamic field.

Seminar July 6–11, 2025 – https://www.dagstuhl.de/25281

2012 ACM Subject Classification Mathematics of computing → Approximation; Mathematics of computing → Computations on polynomials; Mathematics of computing → Quadrature; Mathematics of computing → Interpolation; Applied computing → Engineering

Keywords and phrases exponential analysis, structured matrices, quadrature, subdivision, computer algebra, applications

Digital Object Identifier 10.4230/DagRep.15.7.1

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

From Sparse Interpolation to Signal Processing: New Synergies, *Dagstuhl Reports*, Vol. 15, Issue 7, pp. 1–21
Editors: Annie Cuyt, Dirk de Villiers, Wen-shin Lee, Ana C. Matos, and Gerlind Plonka-Hoch
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Annie Cuyt (University of Antwerp, BE, annie.cuyt@uantwerpen.be)
Dirk de Villiers (University of Stellenbosch, ZA, ddv@sun.ac.za)
Wen-shin Lee (University of Stirling, GB, wen-shin.lee@stir.ac.uk)
Ana C. Matos (University of Lille, FR, ana.matos@univ-lille1.fr)
Gerlind Plonka-Hoch (University of Göttingen, DE, plonka@math.uni-goettingen.de)

In today's data-driven world, where vast volumes of information are generated across scientific, medical, and technological domains, the challenge of extracting meaningful insights from limited, noisy, or high-dimensional measurements has become more pressing than ever. A central problem in this context is the identification of sparse, low-complexity model representations – specifically, models that can accurately describe complex phenomena using the fewest possible parameters or measurements. This is particularly critical in the analysis of multi-exponential signals, where the goal is to recover a small number of exponential components (with real or complex exponents) from a minimal set of observations – often referred to as "probes" or "samples". The acquisition of such measurements is frequently constrained by practical limitations. In medical imaging, for instance, scanning time directly impacts patient comfort and throughput. In remote sensing or radio astronomy, data collection involves expensive instrumentation and limited bandwidth. In industrial testing, measurement costs can be prohibitive. As a result, researchers are increasingly forced to work with datasets that are not only limited in size but also corrupted by noise, making accurate reconstruction a non-trivial task. **This has elevated the need for robust, efficient, and theoretically grounded methods for exponential analysis – methods that can deliver high-fidelity results even under severe data scarcity and noise contamination.** Multi-exponential analysis, though seemingly abstract, plays a surprisingly central role in numerous everyday technologies and scientific disciplines. At its core, it involves decomposing a signal into a sum of exponentials – functions of the form

$$f(t) = \sum_{k=1}^{r} c_k \, \mathrm{e}^{\lambda_k t},$$

where c_k and λ_k may be real or complex.

When the exponents are complex, such models are essential in analyzing oscillatory signals – common in digital signal processing, time series forecasting, and spectral estimation. These techniques are used to extract frequencies, damping rates, and amplitudes from noisy data, forming the backbone of applications ranging from audio and speech processing to financial market modeling and biomedical signal analysis.

When the exponents are real, multi-exponential models describe fundamental physical processes: relaxation dynamics in materials science, decay rates in radioactive isotopes, reaction kinetics in chemistry, heat diffusion in engineering systems, and fluid flow in environmental modeling. These models are not just theoretical constructs – they are indispensable tools for understanding and predicting real-world behavior across physics, biology, and engineering.

The mathematical and computational challenges of multi-exponential analysis are deeply intertwined with several advanced areas of computational science. The problem naturally connects to **structured matrix theory** – where the Hankel or Vandermonde structure

of the data matrix encodes the exponential form – enabling efficient algorithms via rank minimization and low-rank approximation. **Rational approximation theory** provides powerful tools for modeling the Z-transformed signal data as a ratio of polynomials, thereby linking the λ_k and c_k to poles and residues. **Sparse interpolation techniques** allow for the recovery of parameters from few samples, while scale-and-shift invariance principles offer robustness to signal transformations. **Tensor decomposition methods** and **multivariate quadrature** extend these ideas to multi-dimensional data, where the curse of dimensionality poses a fundamental challenge, and advanced techniques including sparsity models are of high importance. Furthermore, **non-convex optimization** plays a crucial role, as the parameter estimation problem often leads to non-convex cost functions with multiple local minima – requiring sophisticated initialization and convergence strategies. **Subdivision methods** further extend the reach of these techniques into geometric and directional data analysis.

The impact of multi-exponential analysis extends far beyond theory. It is foundational in a wide array of **engineering and industrial applications**: direction-of-arrival (DOA) estimation in radar and wireless communications, high-resolution remote sensing and satellite imaging, antenna array design for 5G and beyond, digital image reconstruction and superresolution, precision metrology in manufacturing, radio astronomy for detecting faint cosmic signals, and magnetic resonance imaging (MRI), where fast, accurate reconstruction enables shorter scan times and improved diagnostics. These technologies are not only advancing scientific discovery but are also addressing major societal challenges – improving healthcare outcomes, enabling sustainable energy systems, enhancing transportation safety, supporting space exploration, and strengthening global communication networks.

Given this broad relevance, this Dagstuhl Seminar "From Sparse Interpolation to Signal Processing: New Synergies" (25281) aimed to serve as a vital interdisciplinary forum, bringing together experts from computational harmonic analysis, numerical linear algebra, computer algebra, nonlinear approximation theory, digital signal processing, as well as partners from industry. By fostering dialogue between researchers who have developed similar concepts in isolation, we hope to catalyze cross-fertilization, unify methodologies, and identify shared challenges and opportunities. The goal has been not only to advance the theoretical foundations of exponential analysis but also to accelerate the development of next-generation algorithms that are faster, more robust, and scalable – ultimately enabling breakthroughs in data science, engineering, and beyond.

The talks of this Seminar have been organized with emphasis to the following 6 main topics:

- **Generalisations of exponential analysis**
  (G. Plonka-Hoch, Y. Segman, H. Mhaskar, R. O'Dowd, D. Potts, and J. Prestin),
- **Exponential analysis and structured matrices**
  (A. Matos, H. Liang, M. Ishteva, T. Sauer, and A. Iske)
- **Exponential analysis in computational science**
  (W.-S. Lee, J. Gielis, D. Li, A. Beutler, and R. Beinert)
- **Exponential analysis in quadrature and subdivision**
  (A. Cuyt, T. Perez, M. Cotronei, and M. Piñar)
- **Exponential analysis in engineering**
  (D. de Villiers, D. Davidson, J. Gilmore, N. Diab, and A. Terui)
- **Exponential analysis and computer algebra**
  (J. Gerhard, E. Kaltofen, B. Grenet, and P. Giorgi)

The seminar began on Monday with overview lectures on the first five main topics of the meeting, enabling all participants to quickly gain an entry into the fascinating open interdisciplinary challenges surrounding exponential analysis.

The talk topics on Tuesday covered connections between exponential analysis and sparse approximation, structured matrices, as well as problems in signal analysis and signal separation.

On Wednesday, the topics focused on multivariate integration and subdivision. The free afternoon was used for a short trip to Bernkastel-Kues by several participants and provided the opportunity for physical exercise and lively discussions on the seminar topics.

Thursday was dedicated to various application-oriented topics in exponential analysis. The discussions ranged from sparse models in biology and confocal microscopy to questions concerning the efficient measurement of mutual coupling terms in linear arrays.

Finally, the close connections to special problems in computer algebra, as for example, a quasi-linear time sparse interpolation algorithm over the integers or the fast interpolation and multiplication of unbalanced polynomials, have been the main topic on the final day.

We would like to highlight that this seminar builds upon the 2015 Dagstuhl Seminar 15251, titled "Sparse Modelling and Multi-Exponential Analysis" and the 2022 Dagstuhl Seminar 22221 "Exponential Analysis: Theoretical Progress and Technological Innovation".

The discussions held during the 2015 event sparked numerous fruitful collaborations, including the successful Horizon 2020 RISE project EXPOWER – short for "Exponential Analysis Empowering Innovation" (Grant Agreement No. 101008231, running from 2021-2026), with Annie Cuyt as the coordinator. This project exemplifies how foundational research in exponential analysis can translate into impactful, cross-sectoral innovation.

In October 2025, three months after this meeting, we submitted a new proposal to the Call: Horizon-MSCA-2025-SE-01 (MSCA Staff Exchanges 20225) with the topic TRESUR – "Building synergies between industry and mathematical topics in sparse approximation and recovery", coordinated by Tereza Pérez, in close collaboration with A. Cuyt, W.-S. Lee, A. Matos, M. Piñar, D. de Villiers, G. Plonka-Hoch and several further participants of this Dagstuhl Seminar.

Our experience confirms that Dagstuhl Seminars serve as timely and transformative forums for scientific exchange. They create fertile ground for new partnerships, stimulate interdisciplinary thinking, and unlock novel research directions. In light of rapid advances in both theoretical methods and real-world applications, there is a growing need to strengthen the bridge between cutting-edge mathematical developments and practical industrial challenges. This seminar, and our ongoing efforts, aim to foster such connections – ensuring that theoretical progress continues to inspire and inform real-world innovation.

2 Table of Contents

3 Generalizations of Exponential Analysis

3.1 Prony's Method and Generalizations

Gerlind Plonka (University of Göttingen, DE, plonka@math.uni-goettingen.de)

Joint work of Gerlind Plonka, Kilian Stampfer

The generalized Prony method introduced in [1] is a reconstruction technique for a large variety of sparse signal models that can be represented as sparse expansions into eigen functions of a linear operator A. However, this procedure requires the evaluation of higher powers of the linear operator A that are often expensive to provide. In this survey talk we propose two important extensions of the generalized Prony method that simplify the acquisition of the needed samples essentially and at the same time can improve the numerical stability of the method. The first extension regards the change of operators from A to $\phi(A)$, where ϕ is a suitable operator valued mapping, such that A and $\phi(A)$ possess the same set of eigenfunctions. The goal is now to choose ϕ such that the powers of $\phi(A)$ are much simpler to evaluate than the powers of A. The second extension concerns the choice of the sampling functionals. We show, how new sets of different sampling functionals F_k can be applied with the goal to reduce the needed number of powers of the operator A (resp. $\phi(A)$) in the sampling scheme and to simplify the acquisition process for the recovery method.

This talk is based on joint work with Kilian Stampfer, see [2].

References

1 Peter, T., & Plonka, G. (2013). A generalized Prony method for reconstruction of sparse sums of eigenfunctions of linear operators. Inverse Problems, 29(2), 025001. https://doi.org/10.1088/0266-5611/29/2/025001
2 Stampfer, K., Plonka, G. The Generalized Operator Based Prony Method. Constr Approx 52, 247–282 (2020). https://doi.org/10.1007/s00365-020-09501-6

3.2 Structure Aware Matrix Pencil Method

Yehonatan Segman (Technion – Israel Institute of Technology, Haifa, IL, yehonatans@campus.technion.ac.il)

We address the problem of detecting the number of complex exponentials and estimating their parameters from a noisy signal using the Matrix Pencil (MP) method. We introduce the MP modes and present their informative spectral structure. We show theoretically that these modes can be divided into signal and noise modes, where the signal modes exhibit a perturbed Vandermonde structure. Leveraging this structure, we proposed a new MP algorithm, termed the SAMP algorithm, which has two novel components. First, we present a new and robust model order detection with theoretical guarantees. Second, we present an efficient estimation of signal amplitudes. We show empirically that the SAMP algorithm significantly outperforms the standard MP method, particularly in challenging conditions with closely spaced frequencies and low Signal-to-Noise Ratio (SNR) values. Additionally, compared with prevalent information based criteria, we show that SAMP is more computationally efficient and insensitive to noise distribution.

3.3 Blind Source Signal Separation Using Localized Kernels

Hrushikesh Mhaskar (Claremont Graduate University, US, Hrushikesh.Mhaskar@cgu.edu)

Joint work of Eric Mason, Hrushikesh Mhaskar, Sippanon Kitimoon

The task of separating a superposition of signals into its individual components is a common challenge encountered in various signal processing applications, especially in domains such as audio and radar signals. A previous paper by Chui and Mhaskar [1] proposes a method called Signal Separation Operator (SSO) to find the instantaneous frequencies and amplitudes of such superpositions where both of these change continuously and slowly over time. In this talk, we amplify and modify this method in order to separate linear chirp signals in the presence of crossovers, a very low SNR, and discontinuities.

References

1 Chui, C.K., Mhaskar, H.N. On trigonometric wavelets. Constr. Approx 9, 167–190 (1993). https://doi.org/10.1007/BF01198002

3.4 A Signal Separation View of Classification

Ryan O'Dowd (Claremont Graduate University, US, ryan.o'dowd@cgu.edu)

Joint work of Hrushikesh Mhaskar, Ryan O'Dowd

The main problem of signal separation is to determine the locations ω_k of a signal of the form $\mu(t) = \sum_{k=1}^{K} a_k \, \delta_{\omega_k}(t)$, given observations

$$\widetilde{\mu}(j) = \widehat{\mu}(j) + \varepsilon(j) = \sum_{k=1}^{K} a_k e^{i\omega_k j} + \varepsilon(j), \tag{1}$$

where the $\varepsilon(j)$'s are independent samples from some noise distribution. A key piece of information is the minimal separation among the ω_k's, which dictates the complexity of a model necessary to recuperate the locations to a given accuracy. As this minimal separation tends to zero, we end up in a regime known as super-resolution, which poses its own challenges.

In this work we examine a generalization of the signal separation and super-resolution settings by considering a measure of the form

$$\mu(t) = \sum_{k=1}^{K} c_k \, \mu_k(t),$$

where μ_k's are each measures on some unknown domain. By allowing the constituent measures themselves to be supported on some dense set, our problem of interest incorporates both signal separation and super-resolution as particular cases. We give theory and a method to estimate the supports of the μ_k's given only finitely many samples from μ.

One application of this work lies in machine learning, where we have developed an algorithm for data classification in the active learning paradigm. Therefore, we are able to view signal separation, super-resolution, and machine learning classification problems under a unified umbrella.

3.5 Operator Learning and Sparse Approximation

Daniel Potts (Technische Universität Chemnitz, DE, potts@mathematik.tu-chemnitz.de)

In this talk, we present algorithms for the approximation of multivariate functions. We start with the approximation by trigonometric polynomials based on sampling of multivariate functions on rank-1 lattices. To this end, we study the approximation of functions in periodic Sobolev spaces of dominating mixed smoothness. The proposed algorithm based mainly on a one-dimensional fast Fourier transform, and the arithmetic complexity of the algorithm depends only on the cardinality of the support of the trigonometric polynomial in the frequency domain. After a detailed introduction we will focus on the following questions in more detail.

- We discuss methods where the support of the trigonometric polynomial is unknown.
- We describes an extension of approximation methods for nonperiodic functions via a multivariate change of variables.
- Based on these methods we develop algorithms for discrete operator learning.

References

1 Lutz Kämmerer, Daniel Potts, and Fabian Taubert, Nonlinear approximation in bounded orthonormal product bases. Sampling Theory, Signal Processing, and Data Analysis, 21:19, 2023

2 Daniel Potts and Fabian Taubert, An approach to discrete operator learning based on sparse high-dimensional approximation. arXiv:2406.03973 , 2024

3.6 Sparse Interpolation of Multivariate Functions of Bounded Variation

Jürgen Prestin (University of Lübeck, DE, juergen.prestin@uni-luebeck.de)

Joint work of Jürgen Prestin, E. Semenova

This talk deals with the approximation error of trigonometric interpolation for multivariate functions of bounded variation in the sense of Hardy-Krause. We propose interpolation operators related to both the tensor product and sparse grids on the multivariate torus. For these interpolation processes, we investigate the corresponding error estimates in the Lp norm for the class of functions under consideration. In addition, we compare the accuracy with the cardinality of these grids in both approaches. This is joint work with E. Semenova (Institute of Mathematics, Ukrainian Academy of Sciences, Kiev and University of Lübeck).

4 Structured Matrices

4.1 Structured Matrices and Rational Approximation

Ana Matos (University of Lille, FR, Ana.Matos@univ-lille.fr)

After showing the link between exponential sums models, structured matrices and rational approximation, we will present some results concerning two problems studied within the work package 2 of the EXPOWER project.

The first problem concerns stability problems when leading with Hankel matrix pencils. Given a signal $f(t) = \sum_{j=1}^{r} \alpha_j \exp(\phi_j t)$, with $\alpha_j, \phi_j \in \mathbb{C}$, the aim is to recover the values of the coefficients α_j, $j = 1 \cdots r$ and the (mutually distinct) exponents ϕ_j, $j = 1 \cdots r$. The problem reduces to the computation of eigenvalues of a Hankel pencil,

$$H_n^{(1)} v_j = \lambda_j H_n^{(0)} v_j, \quad H_r^{(m)} = (f_{m+i+j-2})_{i,j=1}^{r}, \quad \lambda_j = \exp(\phi_j \Delta),$$

$$f_k = f(k\Delta),\ k = 0, \cdots, 2r-1$$

where the sampling interval Δ satisfies the Shannon-Nyquist criteria. Starting from a singular pencil $(\widetilde{H}^{(0)}, \widetilde{H}^{(1)})$ polluted by noise of size ϵ, we need to project it into adequate subspaces in order to obtain a regular pencil and then do a perturbation analysis. We obtain upper bounds on the chordal distances between the perturbed and exact eigenvalues and obtain in this way the sensitivity of the eigenvalues. We also get bounds on the euclidean relative error of the corresponding eigenvectors .

The second problem concerns rational approximation and model reduction. From a rational matrix function of type $(N-1, N)$, $H(s) = C(sE - A)^{-1}B + D$, where C, E, A, B are matrices, we are looking for computing recursively strictly proper rational matrix functions H_n of size $m_1 \times m_2$ with Mc-Millan degree $\leq n$, $H_n(s) = C_n(sE_n - A_n)^{-1}B_n + D_n$ satisfying some tangential interpolation conditions. We obtained a formula for the linearized error, and we propose an AAA-type algorithm to compute a sequence of approximants $H_n(s)$ satisfying some tangential interpolation conditions and some error optimization criteria. This is a work in progress.

4.2 Unlabeled Sensing Using Rank-One Moment Matrix Completion

Hao Liang (Chinese Academy of Sciences, Beijing, CN, lianghao2020@amss.ac.cn)

We study the unlabeled sensing problem that aims to solve a linear system of equations $Ax = \pi(y)$ for an unknown permutation π. For a generic matrix A and a generic vector y, we construct a system of polynomial equations whose unique solution satisfies $A\xi^* = \pi(y)$. In particular, ξ^* can be recovered by solving the rank-one moment matrix completion problem. We propose symbolic and numeric algorithms to compute the unique solution. Some numerical experiments are conducted to show the efficiency and robustness of the proposed algorithms.

4.3 Solving Systems of Polynomial Equations with Tensors

Mariya Ishteva (KU Leuven, BE, mariya.ishteva@kuleuven.be)

Joint work of Mariya Ishteva, Philippe Dreesen

Solving systems of polynomial equations is a fundamental problem, with applications in (applied) mathematics, science, and engineering. Although different approaches have been considered in the literature, the problem remains difficult.

Our solution strategy is based on tensor techniques. We first build a partially symmetric tensor from the coefficients of the polynomials. The factors of its (partially symmetric) canonical polyadic decomposition can then be used for building systems of linear equations, which reveal the solutions of the original system.

Future work includes comparisons with existing methods and extending the class of problems, for which the method can be applied.

4.4 Inverses of Multivariate Hankel Matrices

Tomas Sauer (University of Passau, DE & Fraunhofer IIS, DE, Tomas.Sauer@uni-passau.de)

Inverses of Hankel matrices can be given in a somewhat explicit way by means of the so-called Bezoutians. The talk gives the corresponding result for a nonsingular multivariate Hankel matrix, i.e., a matrix formed in the canonical way from a multiindexed moment sequence. It turns out that one obtains a formula for each coordinate direction and that all these formulas involve the orthogonal polynomials as well as the monic H-basis for the associated Prony ideal. This clearly highlights the intimate connection of the problem to exponential polynomials and their reconstruction from equispaced samples.

4.5 A Refined Ingham-Type Theorem for Spectral Properties of Kernel Matrices

Armin Iske (University of Hamburg, DE, armin.iske@uni-hamburg.de)

Joint work of Armin Iske, Tizian Wenzel

We discuss a refined multivariate Ingham-type theorem, whereby we obtain localisation estimates for integrals of exponential sums from symbol functions. This allows us to improve on previous results concerning spectral properties of kernel matrices, including estimates for their spectral condition number. We finally place particular emphasis on spectral alignment for pairs of kernels with finite but different smoothness. This talk is based on joint work with Tizian Wenzel (LMU Munich).

5 Exponential Analysis in Computational Science

5.1 Exponential Analysis Applications

Wen-shin Lee (University of Stirling, GB, wen-shin.lee@stir.ac.uk)

The Blahut/Ben-Or/Tiwari black-box sparse polynomial interpolation algorithm from computer algebra is closely related to exponential analysis, which in turn connects to various areas of computational science. Its links with structured matrix theory, rational approximation and tensor decomposition have opened new possibilities to improve numerical algorithms that are fundamental to a wide range of engineering and industrial applications. These include fluorescence-lifetime imaging microscopy (FLIM), direction-of-arrival (DOA) estimation, remote sensing, antenna design, radar imaging, super-resolution, testing and metrology, radio astronomy, magnetic resonance imaging (MRI), seismology, and financial time series analysis. We report on several recent developments in these areas. Such applications have the potential to address major societal challenges in energy, transportation, space research, healthcare, and telecommunications.

5.2 Inequality and Diversity: Insights from Biology

Johan Gielis (Genicap Beheer BV – Tilburg, NL, johan.gielis@gmail.com)

Inequality is an important topic in economics and human society. In biology similar phenomena are found at the level of individual, organizations and ecosystems. It seems however, that the mathematics underlying these phenomena is very similar, and a new field of research, Econobiology, evolves. Here, we focus on our results from biology and mathematical formulations. The generalized performance equations (product exponentials) are examined in relation to the Lorenz curve, an important tool for measuring income inequality in economics. The relationship between the graphs is provided by 135° rotating and shifted Lorenz curve. This transformation is named Shi Rotations. The results show that the advanced performance models provide an excellent fit for all models tested (leaves in bamboo, melon fruits, tepals of Magnolia flowers and diversity in forests). The Gini coefficient used in economics of inequality is turns out to be closely related to the coefficient of variation, and to other indices such as the Theil index and generalized entropy index. This should lead to new ways of studying nature and human societies, from a dynamical, not a static viewpoint. Although inequality and diversity are two sides of the same coin, this should be done with caution.

References

1 Gielis J. (2024) Performane equations and Shi rotation. Proc. ISSBG2023 Symposium, Geniaal Press, Antwerp, Belgium.
2 Lian, M., Shi P.J., Zhang, L.Y., Yao, W.H., Gielis, J., Niklas, K.J., 2023. A generalized performance equation and its application in measuring the Gini index of leaf size inequality. Trees – Structure and Function, 37:1555–1565.
3 Zhang, L., Quinn, B.K., Hui, C., Lian, M., Gielis, J., Gao, J., Shi, P.J., 2023. New indices to balance α-diversity against tree size inequality. Journal of Forestry Research.

5.3 Compact Non-Invasive Cerebral Blood Flow Sensing

David Li (University of Strathclyde – Glasgow, GB, David.Li@strath.ac.uk)

Continuous, non-invasive monitoring of cerebral blood flow (CBF) is critically important for managing patients in intensive care, particularly neonates and individuals suffering from stroke or traumatic brain injury. Traditional imaging modalities such as MRI or CT remain impractical for bedside or long-term monitoring due to their cost, size, and the need for patient transport. Dr Li's team has developed a compact, low-cost CBF sensing platform based on near-infrared diffuse correlation spectroscopy (DCS), integrated with advanced single-photon avalanche diode (SPAD) sensors. This system offers high sensitivity to microvascular blood flow dynamics with minimal hardware complexity. While current single- or dual-channel systems can monitor CBF changes at discrete sites, the team's vision is to extend this to a multi-channel platform capable of reconstructing three-dimensional images of cerebral perfusion in real time. However, this transition from localised sensing to volumetric imaging poses significant computational and mathematical challenges, especially in addressing the ill-posed inverse problem associated with reconstructing CBF maps from sparse, noisy data. In this presentation, Dr Li will first provide an overview of the physiological motivations for non-invasive, continuous CBF monitoring and outline the limitations of existing technologies. He will then introduce the principles behind DCS and SPAD-based detection, and share results from their current system, including real-time CBF signal recovery in human subjects. The core focus will be on their roadmap towards a multi-channel CBF tomography system, combining hardware innovation with advanced image reconstruction algorithms. The team is particularly keen to engage with mathematicians and computational scientists specialising in inverse problems, compressed sensing, and sparse sampling. Their aim is to identify robust, efficient reconstruction methods to deliver accurate 3D CBF distributions from a minimal number of optical channels, thereby reducing system cost, computational load, and design complexity. By the end of the session, they hope to stimulate interdisciplinary dialogue and explore collaborations to co-develop the next generation of portable, non-invasive cerebral imaging tools that could transform neurocritical care and neonatal monitoring.

5.4 Confocal Microscopy: Different Setups Lead to Different Analysis of the Signals

Andreas Beutler (Mahr GmbH – Göttingen, DE, Andreas.Beutler@mahr.com)

The principle of confocal microscopy is presented. Included is the description of signal structure and challenges of signal analysis. A typical setup has been in practice for many years, however it has limitations for current requirements. We developed a new setup for a much faster and less complicated system which requires a different way of the analysis of the signals. First results are presented.

5.5 Feature Extraction with Applications in Watermark Recognition

Robert Beinert (Technical University of Berlin, DE, robert.beinert@tu-berlin.de)

Joint work of Robert Beinert, Matthias Beckmann, Jonas Bresch

The study of historical watermarks plays a major role in provenance research to determine the date and origin of paper-based writing and art. One of the main watermark collections is the so-called Wasserzeichen-Informationssystem (WZIS) that gathers watermarks from rubbings, handtracings, radiography, and thermography. Since the WZIS consists of nearly as many classes as samples, the training of common deep learning architectures does not yield reliable classifiers. Moreover, digitization techniques like the nowadays employed thermography are highly unstandardized, giving the reason to develop classifies that are invariant under affine image transformations. Based on the so-called Radon cumulative distribution transform (R-CDT), we therefore propose two easy-to-compute feature extractors that facilitate image classification tasks especially in the small data regime and guarantee linear separability of image classes that emerge from affine transformations. Studying the proposed max- and mean-normalized R-CDT, we show robustness against non-affine image deformations. Furthermore, the separability properties of both extractors are stable provided the Wasserstein distance between the samples can be controlled. Our theoretical results are supported by numerical experiments and may pave the path towards computational filigranology.

6 Exponential Analysis in Quadrature and Subdivision

6.1 Exponential Analysis Meets Quadrature and Subdivision

Annie Cuyt (University of Stirling, GB, annie.cuyt@stir.ac.uk)

Exponential analysis is a sparse reconstruction method of an exponential sum of the form

$$f(x) = \sum_{j=1}^{n} \alpha_j \exp(\phi_j x)$$

from $N \geq 2n$ of its equidistantly sampled values $f_s = f(s\Delta), \quad s = 0, \ldots, N-1$, where $\max_j |\Im(\phi_j)\Delta| < \pi$. While the ϕ_j can be obtained via the roots of the well-known formally orthogonal Prony polynomial

$$P_n(u) = \prod_{j=1}^{n} (u - \exp(\phi_j \Delta)),$$

the α_j are the solution of a Vandermonde structured linear system.

The nodes and weights of a Gaussian integration rule follow a similar scheme: the former are the zeroes of some orthogonal polynomial and the latter satisfy a Vandermonde linear system with the given moments f_s on the right hand side.

The annihilation of the values $\exp(\phi_j \Delta)$ by the Prony polynomial is also used in non-stationary subdivision schemes reproducing exponential polynomials, in particular to extract suitable ϕ_j from the data for the reproductive capability.

6.2 From Hermite to Zernike: Orthogonal Polynomials in Optics

Teresa E. Pérez (University of Granada, ES, tperez@ugr.es)

In 1865, Charles Hermite [1] published a paper (divided into four parts) introducing bivariate orthogonal polynomials on the disk to solve a bivariate approximation problem proposed by P. Chebyshev. Despite the fact that a priori the problem seems to be a simple generalization of standard orthogonal polynomials to the bivariate case, the solution presents several obstacles. C. Hermite then introduced the concept of biorthogonality in this context and orthogonal polynomials systems on the disk were described explicitly.

Zernike polynomials were introduced by Frits Zernike en 1934 [2] to describe the wavefront in the formation of images. In 2000, the Optical Society of America (OSA) adopted them as standard patron in Optics and Ophthalmology. Mathematically, Zernike polynomials are polynomials in two variables orthogonal on the unit disk, and are represented in polar coordinates as a product of a radial part (a univariate Jacobi polynomial) and an angular part represented by spherical harmonics.

In this talk we describe the families of bivariate orthogonal polynomials on the disk introduced by C. Hermite, show that Zernike polynomials are a particular case of disk polynomials, and we analyse the main applications of Zernike polynomial in Optics. Finally, our contributions in this topic are presented.

References

1 C. Hermite, Sur quelques d'eveloppement en séries des fonctions, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences. Tome Soixantième. Janvier - Juin 1865. Paris. 370-377, 432-440, 461-466, 512-518.
2 F. Zernike, Beugungstheorie des Schneidenverfahrens und Seiner Verbesserten Form, der Phasenkontrastmethode, Physica. 1 (1934), 689-704.

6.3 Exponential Polynomial Reproduction in Subdivision: Annihilators and Symbol Factorization

Mariantonia Cotronei (Mediterranea University of Reggio Calabria, IT, mariantonia.cotronei@unirc.it)

Subdivision schemes are well-known iterative procedures for generating smooth curves or surfaces from discrete data. A typical requirement for such schemes is their ability to reproduce a function space, that is, to ensure that, starting from sampled data of a continuous function, exactly reconstruct that function in the limit. Focusing on spaces spanned by exponential polynomials, which are crucial in modelling curves of interest in CAGD, such as conics, spirals, or special trigonometric or hyperbolic functions, we explore the close relationship between subdivision schemes and Prony's problem. To do so, we first describe the kernel structure of both convolution and subdivision operators, emphasizing that exponential polynomial sequences are precisely all those that can be annihilated by such operators. We then show how, by making use of such property, the exponential polynomial reproduction capability of the scheme can be fully characterized by factorizing the (level-dependent) subdivision

operator/symbol into specific factors that incorporate the exponential frequencies and their multiplicities. This reveals a strong analogy with Prony's method: while Prony's approach uses annihilating polynomials to recover exponential parameters from sampled data, the subdivision framework uses similar algebraic structures to encode such parameters into the design of the schemes.

6.4 Sobolev Orthogonal Polynomials and Spectral Methods in Boundary Value Problems on the Unit Ball

Miguel Piñar (University of Granada, ES, mpinar@ugr.es)

Our main objective in this work is to demonstrate how orthogonal Sobolev polynomials emerge as a useful tool within the framework of spectral methods for boundary-value problems. The solution of a boundary-value problem for a stationary Schrödinger equation on the unit ball can be studied from a variational perspective. In this variational formulation, a Sobolev inner product naturally arises. As test functions, we consider the linear space of polynomials satisfying the boundary conditions on the sphere, and a basis of mutually orthogonal polynomials with respect to the Sobolev inner product is provided. The basis of the proposed method is provided in terms of spherical harmonics and univariate Sobolev orthogonal polynomials. The connection formula between these orthogonal Sobolev polynomials and classical orthogonal polynomials on the ball is established. Consequently, the Sobolev Fourier coefficients of a function satisfying the boundary value problem are recursively derived. Finally, numerical experiments were presented.

7 Exponential Analysis in Engineering

7.1 Sparsity in Antenna Engineering

Dirk de Villiers (Stellenbosch University, ZA, ddv@sun.ac.za)

For the last several years a fruitful collaboration with the EXPOWER H2020 RISE project, and specifically Antwerp University, resulted in the application of various forms of sparse exponential analysis on antenna engineering problems. The talk presents several such example applications including:

- Direction of Arrival Estimation using 1-bit sampled data
- Near-field antenna position estimation using drone measurements
- Frequency ripple characterisation in reflector antenna noise temperate calculations
- Progress towards compact field pattern storage

7.2 Mathematical Challenges for Low-Frequency Radio Telescope Design

David B Davidson (ICRAR-Curtin, Perth, AU, David.Davidson@curtin.edu.au)

The overall theme of this paper is outstanding mathematical challenges in modelling (and designing) low-frequency (30-300 MHz) radio telescopes, epitomised by SKA-Low, the low-frequency component of the Square Kilometre Array radio telescope.

The presentation starts with a review of interferometric radio telescopes, and then moves on to introduce aperture arrays, which are receive-only phased arrays in conventional antenna terminology. Following this, Computational Electromagnetic (CEM) techniques [1] for simulating radio telescopes are addressed. The Method of Moments (MoM) has proven especially useful for wire antenna arrays, as widely used in low-frequency telescopes, and when combined with the Multi-Level Fast Multipole Method (MLFMM) acceleration technique, permits the analysis of large arrays, such as the 256-antenna SKA-Low "station" [2]. Nonetheless, these remain formidably large problems, with several million unknowns. Typical run-times for an SKA-Low station take hours to days for each frequency, depending on the rate of convergence of the MLFMM, even using high-performance computing resources.

Mutual coupling and its impact on telescope performance is a major theme of the talk. It is shown that the effect of mutual coupling is largely negative. The effects can be predicted but only by using a full simulation as above. In terms of SKA-Low science which may be impacted by this, the Epoch of Reionisation (EOR) is one of its major science cases. This requires looking for a signal five orders of magnitude in power below the foreground signals. The power spectrum approach is outlined, which uses Fourier analysis to attempt to distinguish the smooth foreground signal from the desired EOR signal. Recent work has demonstrated that mutual coupling poses major challenges to this approach.

The paper concludes with an outline of how new mathematical methods could assist in designing future low-frequency radio telescopes, combined with studies of more modest antenna systems with less overall complexity for initial work. Tools such as surrogate modelling could be very useful, but optimisation goals will need careful thought.

References

1 D.B. Davidson, "Computational Electromagnetics for RF and Microwave Engineering", 2nd ed, Cambridge University Press, Cambridge, 2011.

2 P. Bolli, D. B. Davidson, M. Bercigli, P. Di Ninni, M. G. Labate, D. Ung, and G. Virone, "Computational electromagnetics for the SKA-Low prototype station AAVS2," Journal of Astronomical Telescopes, Instruments,and S ystems, vol. 8, no. 1, p. 011017, 2022, doi: 10.1117/1.JATIS.8.1.011017.

7.3 Measuring Linear Array Mutual Coupling Terms using Exponential Analysis

Jacki Gilmore (Stellenbosch University, ZA, jackivdm@sun.ac.za)

We present a measurement-based method for jointly estimating the element positions and mutual-coupling coefficients of uniform linear arrays (ULAs) directly from embedded-element radiation patterns. The problem is cast as a Prony model whose shared exponential bases encode the electrical spacing (frequency and element locations), while the model coefficients capture the mutual coupling. By applying multi-snapshot validated exponential analysis (VEXPA) to the noisy pattern samples, the common bases are retrieved with sufficient resilience to noise. The coefficients are then obtained from an overdetermined Vandermonde system that is nearly perfectly conditioned. The technique is demonstrated on two cases:

- Synthetic example – An 11-element dipole ULA spanning 800–1200 MHz, sampled at 51 frequencies, with added white Gaussian noise (SNR = 60 dB). Using 361 pattern samples, the base terms and coefficients are both recovered. The relative error of the coefficients is in the order of 10-5,
- Measured array – A 4-element dipole ULA measured from 2.8–3.2 GHz at nine frequencies (again 361 samples). Both the base terms and the coefficients were extracted, allowing the mutual coupling terms to be determined with sufficient accuracy.

7.4 Selecting Sampling Rates and Sets for Efficient Super Resolution

Nuha Diab (Tel-Aviv University, IL, nuhadiab@tauex.tau.ac.il)

In the first part of the talk, we investigate the recovery of nodes and amplitudes from noisy frequency samples in spike train signals, also known as the super-resolution (SR) problem. When the node separation falls below the Rayleigh limit, the problem becomes ill-conditioned. Admissible sampling rates, or decimation parameters, improve the conditioning of the SR problem, enabling more accurate recovery. We propose an efficient preprocessing method to identify the optimal sampling rate, significantly enhancing the performance of SR techniques.

For the second part of the talk, we study the spectral properties of infinitely smooth multivariate kernel matrices when the nodes form a single cluster. We show that the geometry of the nodes plays an important role in the scaling of the eigenvalues of these kernel matrices. For the multivariate Dirichlet kernel matrix, we establish a criterion for the sampling set ensuring precise scaling of eigenvalues. Additionally, we identify specific sampling sets that satisfy this criterion. Finally, we discuss the implications of these results for the problem of super-resolution, i.e. stable recovery of sparse measures from bandlimited Fourier measurements.

7.5 Solving Estimation Problems Using Minimax Polynomials and Gröbner Bases

Akira Terui (University of Tsukuba, JP, terui@math.tsukuba.ac.jp)

An estimation problem is a problem in which one infers or estimates a certain quantity or state based on uncertain information or observed data. Estimation problems play a crucial role across various disciplines, including statistics, machine learning, signal processing, and control theory. To solve estimation problems, numerical methods such as the gradient method or genetic algorithms are used. However, gradient methods may return a local solution, depending on the initial values, since they utilize local convergence properties. The genetic algorithm has some disadvantages, as it sometimes fails to properly solve the estimation problem due to phenomena such as initial convergence and hitchhiking. On the other hand, using minimax approximation together with Gröbner bases computation may avoid these phenomena, for this method evaluates values globally. In this presentation, we propose a method for solving estimation problems using minimax polynomials and Gröbner bases. We show an application of the proposed method for solving a speech direction estimation problem.

8 Computer Algebra

8.1 What's New in Maple 2025

Jürgen Gerhard (Maplesoft – Waterloo, CA, jgerhard@maplesoft.com)

We will highlight some of the new features in Maple 2025, including for advanced mathematical computations, user interface redesign, graph theory, visual expression comparison, code generation, and programming.

8.2 Sparse Interpolation in Chebyshev Basis: Early Termination and Georg Heinig's Toeplitz Solver

Erich Kaltofen (North Carolina State University – Raleigh, US, kaltofen@ncsu.edu)

Joint work of Erich Kaltofen, Zhi-Hong Yang

Ideas by Kaltofen and Yang [1] for error-correcting interpolation of polynomials that are a sparse linear combination of Chebyshev polynomials have led to a new early termination algorithm for computing the sparsity.

Kaltofen and Lee [2] in their early termination algorithms used thresholds to skip over sporadic probabilistic errors. For early termination in sparse Chebyshev interpolation, thresholds need an algorithm to step from a sequence of singular leading principal submatrices of a Toeplitz matrix to the next non-singular leading principal submatrix. For Prony sparse interpolation, the problem is solved by the 1969 Berlekamp-Massey algorithm, and for Chebyshev sparse interpolation by Georg Heinig's 1983 Toeplitz algorithm.

In my talk, I will describe our new early termination algorithm and Heinig's Toeplitz solver from a Berlekamp-Massey algorithmic viewpoint. Heinig's algorithm, which generalizes the classical Toeplitz solvers by Levinson and Durbin, takes quadratic time and requires linear space.

This is joint work with Zhi-Hong Yang at Central South University, China.

References

1 Erich L. Kaltofen and Zhi-Hong Yang. Sparse Polynomial Interpolation With Error Correction: Higher Error Capacity by Randomization. In Proceedings of the International Symposium on Symbolic and Algebraic Computation (ISSAC '24), July 16–19 2024, Raleigh, NC, USA. ACM, 2024. DOI: 10.1145/3666000.3669698.
2 Kaltofen, E. L. & Lee, W.-s. "Early termination in sparse interpolation algorithms." Journal of Symbolic Computation, Vol. 36 (3–4), 2003, pp. 365–400.

8.3 Quasi-Linear Interpolation of Sparse Polynomials Over the Integers

Bruno Grenet (LIRMM, University of Montpellier, CNRS Montpellier, FR, bruno.grenet@lirmm.fr)

Joint work of Bruno Grenet, Pascal Giorgi, Armelle Perret du Cray, Daniel S. Roche

Prony's method can be used for sparse interpolation over any ring: Given black box access to a polynomial with t non-zero terms, its coefficients and exponents can be computed from evaluations on a geometric sequence of size $2t$. However, over exact rings such as the integers or finite fields, this algorithm is computationally expensive. Several techniques have been developed by the computer algebra community to speed up the algorithm. As a result, a sparse polynomial over the integers can be interpolated at a cost that is quasi-linear in the size of its sparse representation.

8.4 Fast Interpolation and Multiplication of Unbalanced Polynomials

Pascal Giorgi (LIRMM, University of Montpellier, CNRS Montpellier, FR, pascal.giorgi@lirmm.fr)

Joint work of Pascal Giorgi, Bruno Grenet, Armelle Perret du Cray, Daniel S. Roche

Efficient polynomial or integer multiplication is at the core of computer algebra and these problems received a lot of attention since the last past decades to reach quasi-linear time complexity. Nowadays, it is even possible to multiply polynomials with integer coefficients within a quasi-optimal complexity. Note that the latter result assumes that the bit-lengths of the coefficients must not vary too much, meaning that polynomials with very unbalanced coefficients are out of reach yet. In this talk, I will show how this problem of unbalanced integer polynomials multiplication is related to sparse polynomial interpolation. By using our recent technique on sparse interpolation for integer polynomial, we show how we can reach an almost quasi-optimal complexity for the multiplication problem. In particular, we will describe a new algorithm that enables the interpolation of sparse unbalanced polynomials in almost quasi-linear time.

Participants

- Bernhard Beckermann
University of Lille, FR
- Robert Beinert
TU Berlin, DE
- Andreas Beutler
Mahr – Göttingen, DE
- Mariantonia Cotronei
University Mediterranea of Reggio Calabria, IT
- Annie Cuyt
University of Antwerp, BE
- David Davidson
Curtin University – Bentley, AU
- Dirk de Villiers
Stellenbosch University, ZA
- Nuha Diab
Tel Aviv University, IL
- Jürgen Gerhard
Maplesoft – Waterloo, CA
- Johan Gielis
Genicap – Tilburg, NL
- Mark Giesbrecht
University of Waterloo, CA
- Jacki Gilmore
Stellenbosch University, ZA
- Pascal Giorgi
University of Montpellier & CNRS, FR
- Bruno Grenet
University of Grenoble, FR
- Mariya Ishteva
KU Leuven – Geel, BE
- Armin Iske
Universität Hamburg, DE
- George Labahn
University of Waterloo, CA
- Wen-shin Lee
University of Stirling, GB
- David Li
The University of Strathclyde – Glasgow, GB
- Hao Liang
Chinese Academy of Sciences – Beijing, CN
- Ana C. Matos
Lille I University, FR
- Hrushikesh N. Mhaskar
Claremont Graduate University, US
- Hans Michael Möller
TU Dortmund, DE
- Ryan O'Dowd
Claremont Graduate University, US
- Anthony O'Hare
University of Stirling, GB
- Miao-Jung Yvonne Ou
University of Delaware, US
- Teresa E. Pérez
University of Granada, ES
- Miguel Piñar
University of Granada, ES
- Petr Plechac
University of Delaware, US
- Gerlind Plonka-Hoch
Universität Göttingen, DE
- Daniel Potts
TU Chemnitz, DE
- Jürgen Prestin
Universität zu Lübeck, DE
- Michele Pugno
University of Antwerp, BE
- Daniel Roche
U.S. Naval Academy – Annapolis, US
- Tomas Sauer
Universität Passau, DE
- Yehonatan-Itay Segman
Technion – Haifa, IL
- Ramonika Sengupta
TU Eindhoven, NL
- Richard G. Spencer
National Institutes of Health – Baltimore, US
- Akira Terui
University of Tsukuba, JP
- Lihong Zhi
MMRC – Beijing, CN

Remote Participants

- Erich Kaltofen
North Carolina State University – Raleigh, US

Report from Dagstuhl Seminar 25282

Theory of Neural Language Models

Pablo Barcelo[*1], David Chiang[*2], George Cybenko[*3], Lena Strobl[*4], and Andy Yang[†5]

1 **PUC – Santiago de Chile, CL.** pbarcelo@ing.puc.cl
2 **University of Notre Dame, US.** dchiang@nd.edu
3 **Dartmouth College Hanover, US.** george.cybenko@dartmouth.edu
4 **University of Umeå, SE.** lena.strobl@gmail.com
5 **University of Notre Dame, US.** ayang4@nd.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25282 "Theory of Neural Language Models". The seminar aimed to bring researchers together to lay a foundation for continued work on the theory of neural language models, focusing on questions including: How do transformers, RNNs, other NLMs, and their variants, compare with one another in expressivity and trainability? How do the successes and failures of NLMs predicted by theoretical models manifest in practice? What modifications, or what wholly new architectures, are suggested by the theory?

Seminar July 06–11, 2025 – https://www.dagstuhl.de/25282
2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Theory of computation → Formal languages and automata theory; Theory of computation → Logic; Theory of computation → Circuit complexity; Theory of computation → Communication complexity
Keywords and phrases Dagstuhl Seminar, Neural Networks, Language Models, Automata, Logic, Model Theory, Circuit Complexity
Digital Object Identifier 10.4230/DagRep.15.7.22

1 Executive Summary

Pablo Barcelo (PUC – Santiago de Chile, CL)
David Chiang (University of Notre Dame, US)
George Cybenko (Dartmouth College Hanover, US)
Lena Strobl (University of Umeå, SE)

Artificial intelligence (AI) has gone through multiple "summers" and "winters," with the current summer based on large neural models that generate text, images, and other content. ChatGPT and other neural language models (NLMs), which model sequences of tokens, have taken center stage, not only in natural language processing, but across a wide range of applications. This interest spans the academic, corporate, government, investor, and consumer sectors. However, whereas experimental research and product development are surging ahead, more theoretical research aimed at foundational questions about neural networks is lagging behind. Clear thinking about what NLMs can and can't do is more needed than ever.

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

Theory of Neural Language Models, *Dagstuhl Reports*, Vol. 15, Issue 7, pp. 22–52
Editors: Pablo Barcelo, David Chiang, George Cybenko, Lena Strobl, and Andy Yang
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The old guard of NLMs, recurrent neural networks (RNNs), have been studied theoretically for decades, in relation to finite automata and Turing machines. At present, the dominant NLMs are based on transformers, whose computational power is a new and rapidly growing area of research. Transformers have been related to a wide variety of formal models from computability and complexity theory, like counter automata, Turing machines, Boolean circuits, and first-order logic. However, a unified and comprehensive theory of the abilities and limitations of transformers is not yet in sight.

Such a theory would ideally answer questions like:

- How do transformers, RNNs, other NLMs, and their variants, compare with one another in expressivity and trainability?
- How do the successes and failures of NLMs predicted by theoretical models manifest in practice?
- What modifications, or what wholly new architectures, are suggested by the theory?

There is a small but growing community of researchers that investigates such questions about NLMs. This Dagstuhl Seminar aimed to bring this community together to lay a foundation for continued work in this area, identifying central open problems, and fostering new collaborations. To achieve these goals, the seminar left ample room for informal discussions on topics suggested by the participants themselves, along the lines of an Open Space Technology meeting.

2 Table of Contents

3 Tutorials

3.1 Transformers

David Chiang (University of Notre Dame, US)

On Day 1, I gave an introduction to transformers, going over the basic definition briefly. (Andy Yang later gave a more detailed introduction for participants coming from outside AI.) Then I gave an overview of results proving exact equivalence between variants of transformers and various complexity classes, and focused on three themes that I thought would provoke discussion:

- The equivalence results naturally fall into two groups: on the one hand, equivalences between transformers and $\mathsf{FO}[<]$-uniform complexity classes; and on the other hand, equivalences between transformers *with intermediate steps* and DLOGTIME-uniform ($= \mathsf{FO}[+, \times]$-uniform) complexity classes. (Thanks to Michaël Cadilhac for this formulation.) The use of intermediate steps, then, is a particularly important variation.
- Another important variation is in *uniformity* or aspects of the network depending on n, the input length. I distinguished between "parameter-uniformity," where the parameters do not depend on n but other aspects might, and setups where the parameters may also depend on n.
- A very large number of variations can be seen as workarounds for a result by [1]: If a transformer uses softmax attention, has bounded position embeddings, and has only Lipschitz-continuous position-wise functions, then a change in a single input symbol results in an $O(1/n)$ change in the output, where n is the input length. If, furthermore, a minimum gap is required between different outputs (e.g., accept vs. reject), then a transformer can only solve trivial problems. So, many constructions use tricks to work around one of the above assumptions.

This overview was based in part on the survey by [2].

References

1 Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306. URL https://aclanthology.org/2020.tacl-1.11.

2 Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages can transformers express? A survey. *Transactions of the Association for Computational Linguistics*, 12:543–561, 2024. doi: 10.1162/tacl_a_00663.

3.2 Theoretical Background

Howard Straubing (Boston College, US)

I discussed classes of formal languages – some classes of regular languages, classes defined by different kinds of logical formulas (variants of first order and temporal logic), and circuit complexity classes – that have been studied for many years by theoretical computer scientists, and that are receiving renewed attention because of their connection to the expressive power of transformers.

Regular languages:

$$\begin{aligned}\mathsf{FO}[<] &= \text{star free} = \mathsf{LTL} \text{ (linear temporal logic)}\\ &\subsetneq \mathsf{FO}[<, \mathsf{MOD}] \text{ (modular predicates)}\\ &\subsetneq \mathsf{FO{+}MOD}[<] \text{ (modular quantifiers)}\\ &\subsetneq \text{regular languages.}\end{aligned}$$

Circuit complexity classes:

$$\begin{aligned}\mathsf{AC}^0 &= \mathsf{FO}[\mathsf{Any}]\\ &\subsetneq \mathsf{ACC} = \mathsf{FO{+}MOD}[\mathsf{Any}]\\ &\subsetneq^{?} \mathsf{TC}^0 = \mathsf{Maj}[\mathsf{Any}] \text{ (majority quantifiers)}\\ &\subsetneq^{?} \mathsf{NC}^1\end{aligned}$$

where $\subsetneq^{?}$ means that inclusion is known, but strictness is not known.

I also discussed the uniform versions of the circuit complexity classes. Questions of inclusion among these classes are addressed by algebraic means, especially via the syntactic monoids of regular languages.

3.3 Methods and Architectures to Increase Expressivity

William Merrill (New York University, US)

We have seen how transformers (when used as classifiers or next-token predictors) cannot express inherently sequential computation outside the class TC^0. In contrast, many applications of LLMs, e.g., state tracking or reasoning about math or code, conceivably require sequential processing. This tutorial covered two ways to extend transformers' expressive power so they can solve inherently sequential computational problems:

1. Extending transformers with **chain of thought** allows them to simulate Turing machine steps, enabling computation outside TC^0 with enough chain-of-thought steps. However, this sacrifices parallelism in order to gain sequential expressive power.
2. Moving away from transformers, **linear RNNs** (a.k.a. state-space models) can gain some expressive power outside TC^0 while maintaining moderate parallelism on current hardware.

3.3.1 Chain of thought (CoT)

We view a transformer as a function $T : \Sigma^* \to \Sigma$ by taking the argmax logits. Without CoT, we solve a decision problem with T by simply computing $T(w)$ on input w and interpreting specific tokens as yes/no. With CoT, we allow T to autoregressively generate t "thinking" tokens before generating a final token $t + 1$ as yes/no output.

So, the full context generated by CoT looks like:

$$\underbrace{x_1 \dots x_n}_{\text{Input tokens}} \quad \underbrace{z_1 \dots z_t}_{\text{CoT tokens}} \quad \underbrace{y}_{\text{Output}}$$

▶ **Definition 1.** For any function $t(n)$, let $\mathsf{CoT}[t]$ be the class of problems expressible by a transformer that is allowed $O(t(n))$ steps of CoT on inputs of length n.

For example, $\mathsf{CoT}[\log n]$ represents *logarithmic CoT* and $\mathsf{CoT}[n]$ represents *linear CoT*. We will also refer to $\cup_{c=0}^{\infty}\mathsf{CoT}[n^c]$ as *polynomial CoT*.

3.3.1.1 Simulating Turing machines with CoT

Before seeing how transformers can use CoT to simulate Turing machines, we briefly review multitape Turing machines:

- Finite state, read-only input tape, and several work tapes
- At step i, each tape τ has contents Γ_i^τ and head position h_i^τ
- The Turing machine is specified a transition function δ that reads the current symbol on each tape head as well as some finite state and returns a new state and new output symbols and move directions on each tape. That is, for states $q, q' \in Q$, tape symbols $\gamma^\tau, \gamma'^\tau \in \Gamma$, and move directions $m^\gamma \in \{-1, 0, 1\}$,

$$\delta : q, \gamma^0, \gamma^1, \ldots \gamma^k \mapsto q', \gamma'^0, \gamma'^1, \ldots \gamma'^k, m^0, m^1, \ldots, m^k$$

- Having computed δ, we update the finite state, tape contents, and head position:

$$\begin{aligned} q_{i+1} &= q'_i \\ \Gamma^\tau_{i+1}[h^\tau_i] &= \gamma'^\tau_i \\ h^\tau_{i+1} &= h^\tau_i + m^\tau_i \end{aligned}$$

▶ **Theorem 2** ([1, 2])**.** *CoT transformers can simulate multitape Turing machines. That is, for any $t(n)$, $\mathsf{TIME}[t] \subseteq \mathsf{CoT}[t]$.*

Proof sketch. The idea will be to maintain an immutable representation of the Turing machine tape distributed across CoT tokens. Each token will store the symbols written to the tape at the previous step, as well as a pointer to the current head position. With this information, future CoT tokens can always attend back to reconstruct the last symbol written to a particular position.

We proceed by induction, showing that CoT token i can encode q_i, γ_i^τ (the token written by step $i-1$), and m_i (the move after step $i-1$).

1. **Compute Head Pointer:** At each token i, we will recover h_i^τ, the head position on tape τ at step i. This can be done by counting the number of right and left moves on each tape, i.e., as $\sum_{j=1}^{i} m_j$.
2. **Retrieve Tape Symbol:** Once we have h_i^τ, we recover the current symbol under head τ via an "induction head" [3] construction: we will find the last token where tape τ was at h_i^τ and retrieve the token outputted immediately after that. To do this, the head will hard attend from step i to the last step j such that $h_i^\tau = h_{j-1}^\tau$ and retrieve the CoT token δ_j.
   In more detail, one way to implement this head with saturated attention is as follows. Adapting [2], let $\phi(x)$ be the projection of $\langle x, 1\rangle$ onto the unit sphere and let $f(i)$ be small and monotonically decreasing.
   - **Query i:** $\langle \phi(h_i^\tau), -1\rangle$
   - **Key j:** $\langle \phi(h_{j-1}^\tau, 1), \phi(f(i))\rangle$
   - **Value j:** γ_j^τ, i.e., the token written to h_{j-1}^τ

This head returns γ_j^τ from the largest j such that $h_i^\tau = h_{j-1}^\tau$, i.e., the current symbol on tape τ at position h_i^τ.

3. **Compute Transition Function:** Once we have computed the symbol γ_i^τ on each tape, we can use a feedforward network to compute the following transition function via a finite lookup table:

   $$\delta(q_i, \gamma_i^0, \gamma_i^1, \ldots, \gamma_i^k).$$

   We then output a single CoT token that encodes the symbols that should be written and directions that should be moved – this means the vocab size increases with the number of state and tapes but is fixed w.r.t. n.

We conclude that we can simulate a Turing machine for t steps using t CoT tokens. ◀

Assuming $t \leq \text{poly}(n)$, this construction only requires $O(\log n)$ precision, since the growing tape is distributed across different tokens. This is different than the classical RNN Turing completeness construction for RNNs [4], which requires the precision to grow linearly with runtime.

3.3.1.2 Understanding how much CoT is required for more expressive power

We can consider the expressive power of transformers with different amounts of CoT. Immediately, we see that polynomial CoT allows transformers to solve any problem solvable in polynomial time:

▶ **Corollary 3.** $\bigcup_{c=0}^{\infty} \mathsf{CoT}[n^c] = \mathsf{P}$.

But this is a lot of CoT. Do we still gain expressive power with shorter CoT lengths? Yes, with less CoT, we still can solve some problems likely outside TC^0:

▶ **Corollary 4** ([2, 5])**.** $\mathsf{CoT}[n]$ *contains* NC^1*-complete problems including regular language recognition and boolean formula evaluation.*

What about sublinear CoT? Interestingly, logarithmic CoT remains in TC^0:

▶ **Theorem 5** ([6])**.** $\mathsf{CoT}[\log n] \subseteq \mathsf{TC}^0$.

Proof sketch. We aim to simulate a transformer that uses $c \log n$ CoT steps with a TC^0 circuit family. To do this, we enumerate all possible CoT's using $O(n^c)$ constant gates. In parallel for each token of each possible CoT, we apply a TC^0 circuit (corresponding to the transformer) that checks whether the current token would have been produced by its left context. We select the unique CoT where each token matches the transformer's output. ◀

So more than logarithmic CoT is required to gain expressive power outside TC^0.

An interesting open question is proving lower bounds on CoT length against transformers. [7] make exciting progress on this assuming hard-attention transformers.

3.3.1.3 Alternatives to CoT: padding and looping

Padded transformers: augment the transformer with a CoT of blank tokens ($\square$) rather than model-generated tokens. Unlike standard CoT, this is parallelizable because we don't need to wait for the output of the first t tokens before processing token $t + 1$.

▶ **Theorem 6** ([8, 9, 10])**.** *Soft-attention transformers with polynomial padding tokens express exactly* TC^0.

Looped transformers: rather than autoregressively generating tokens, just repeat layers in the transformer. Also called "universal transformers".

▶ **Theorem 7** ([11, 12]). *Looped transformers can solve the* NC^1*-complete problem of regular language recognition.*

Moreover, transformers with polylogarithic looping and polynomial padding can express all of NC, showing that these methods can unlock substantial expressivity while preserving moderate parallelism (unlike CoT).

3.3.1.4 Takeaways about CoT and related methods

- CoT gains expressivity but sacrifices parallelism and requires many steps!
- Padding is parallelizable but doesn't gain expressivity outside TC^0
- Looping gains expressivity even with a small number of steps: relatively parallelizable
- Parallelism tradeoff: expressivity for sequential problems and parallelism are at odds

3.3.2 New architectures: linear RNNs

Rather than relying on CoT for greater expressive power, can we design new architectures that increase expressivity while maintaining parallelism? To get sequential expressivity, we could look back to RNNs:

▶ **Definition 8** (Nonlinear RNN)**.** For some nonlinearity σ, an RNN has the form

$$h_{i+1} = \sigma(Ah_i + Bx_i).$$

These RNNs can represent the NC^1-complete problem of recognizing regular languages [13]. But the nonlinearity means that it's not easy to parallelize RNNs.

A new line of work has therefore turned to linear RNNs (also called state-space models):

▶ **Definition 9** (Linear RNN)**.**

$$h_{i+1} = A_i h_i + Bx_i.$$

Relative to old RNNs, linear RNNs are parallelizable. Relative to transformers, they use less memory have the potential for greater expressivity on sequential problems.

S4 and Mamba are particular instantiations of linear RNNs (there are many others). Linear attention is also related:

▶ **Definition 10** (Linear attention)**.**

$$h_i^\top = \sum_{j=1}^{i} q_i^\top k_j \cdot v_j^\top.$$

Linear attention can be written as the following linear RNN:

$$\begin{aligned} S_{i+1} &= S_i + k_j v_j^\top \\ h_i^\top &= q_i^\top S_i. \end{aligned}$$

3.3.2.1 Parallelizing linear RNNs

Leveraging linearity, we can rewrite a linear RNN in its "convolutional form":

$$h_{i+1} = A_i h_i + Bx_i$$
$$h_i = \sum_{j=1}^{i} A_i A_{i-1} \dots A_{j+1} B x_j.$$

This can be computed efficiently as a prefix sum in the monoid generated by $\langle A_i, Bx_i \rangle$ where $\oplus$ is defined[1]

$$\langle A_1, b_1 \rangle \oplus \langle A_2, b_2 \rangle = \langle A_2 A_1, A_2 b_1 + b_2 \rangle.$$

Using efficient algorithms for prefix sums [14], we can compute this efficiently.

3.3.2.2 Expressivity benefits of linear RNNs

Does the recurrent nature of linear RNNs enable them to represent NC^1-complete problems like regular language recognition? For several popular early instantiations of linear RNNs, the answer is no, but, in general, it is possible.

▶ **Theorem 11** ([15]). *If $A_i = A$ (e.g., in S4), then a linear RNN is in TC^0.*

Proof. Computing the linear RNN reduces to

$$h_i = \sum_{j=1}^{i} A^{i-j} B x_j.$$

Iterated addition and matrix powering are in TC^0, so this can be computed in TC^0. ◀

▶ **Theorem 12** (Informal; [15]). *If A_i is diagonal and parameterized reasonably (e.g., in Mamba), then a linear RNN is in TC^0.*

Proof sketch. Since A_i's are diagonal, computing the product $A_i A_{i-1} \dots A_{j+1}$ reduces to iterated multiplication of scalars in parallel, which is commutative and in TC^0. ◀

However, it's possible to make linear RNNs expressive enough for regular language recognition by allowing A_i to be more complex:

▶ **Theorem 13** ([15]). *Let A_i be a learned embedding of token σ_i (call this IDS4). For any regular language, there exists an IDS4 linear RNN that recognizes it.*

Proof sketch. We can directly encode FSA transitions with such a linear RNN. We assume a beginning of sequence symbol \$. We set B\$ to return a representation of the initial state (a one-hot diagonal matrix). We use A_i to represent the transition monoid element δ_i associated with σ_i. The state is thus

$$h_i = (\delta_i \circ \delta_{i-1} \circ \dots \circ \delta_1)(q_0).$$

We can then use a final linear layer to detect whether h_i represents an accepting state. ◀

[1] Nice presentation here: `https://openreview.net/pdf?id=RDbuSCWhad`

▶ **Corollary 14.** *IDS4 can solve an* NC^1*-complete problem.*

Downside: letting A_i be an arbitrary matrix requires $O(d^2)$ memory instead of $O(d)$ for a diagonal matrix.

Recent work has come up with new A_i parameterizations that use less (i.e., linear) memory but still can recognize regular languages. Some of these linear RNNs have been scaled up recently with encouraging signs!

- DeltaNet++ [16]: an extension to DeltaNet linear RNN similar to IDS4 unlocks greater expressivity and improves math/code reasoning at the 7B scale
- RWKV-7 [17]: 7B linear RNN that can express regular language recognition and achieves strong LM performance
- PaTH attention [18]: an extension to transformers inspired by DeltaNet extension that increases expressivity, synthetic evaluations, length generalization, and language modeling performance

3.3.2.3 Expressivity drawbacks of linear RNNs

Major drawback of linear RNNs relative to transformers: they cannot copy or retrieve arbitrary tokens from the past due to their bounded memory [19]

- As an implication, they cannot solve associative recall [20] or form induction heads [3], which have been shown to be important for language modeling
- The Turing machine simulation construction in Theorem 2 utilizes induction heads, so linear RNNs can't implement it

Potential solutions:

- Mix linear RNN layers with at least one transformer layer for retrieval
- Augment linear RNNs with some other retrieval mechanism

References

1. Jorge Pérez, Pablo Barceló, and Javier Marinkovic. Attention is Turing-complete. *Journal of Machine Learning Research*, 22(75):1–35, 2021. URL `http://jmlr.org/papers/v22/20-302.html`.
2. William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NjNGlPh8Wh`.
3. Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL `https://arxiv.org/abs/2209.11895`. arXiv:2209.11895.
4. H.T. Siegelmann and E.D. Sontag. On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1):132–150, 1995. doi: `10.1006/jcss.1995.1013`. URL `https://www.sciencedirect.com/science/article/pii/S0022000085710136`.
5. Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind Chain of Thought: A theoretical perspective. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 70757–70798, 2023. URL `https://papers.nips.cc/paper_files/paper/2023/hash/dfc310e81992d2e4cedc09ac47eff13e-Abstract-Conference.html`.

6 Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3EWTEy9MTM.

7 Alireza Amiri, Xinting Huang, Mark Rofin, and Michael Hahn. Lower bounds for chain-of-thought reasoning in hard-attention transformers. In *Proceedings of ICML*, 2025. URL https://arxiv.org/abs/2502.02393.

8 Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=NikbrdtYvG.

9 William Merrill and Ashish Sabharwal. Exact expressive power of transformers with padding, 2025a. URL https://arxiv.org/abs/2505.18948. arXiv:2505.18948.

10 Charles London and Varun Kanade. Pause tokens strictly increase the expressivity of constant-depth transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2505.21024.

11 Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=De4FYqjFueZ.

12 William Merrill and Ashish Sabharwal. A little depth goes a long way: The expressive power of log-depth transformers, 2025b. URL https://arxiv.org/abs/2503.03961. arXiv:2503.03961.

13 William Merrill. Sequential neural networks as automata. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pages 1–13, August 2019. doi: 10.18653/v1/W19-3901. URL https://aclanthology.org/W19-3901/.

14 Guy E. Blelloch. Prefix sums and their applications. Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University, November 1990.

15 William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=QZgo9JZpLq.

16 Riccardo Grazzi, Julien Siems, Jörg K.H. Franke, Arber Zela, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UvTo3tVBk2.

17 Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S. Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. RWKV-7 "Goose" with expressive dynamic state evolution, 2025. URL https://arxiv.org/abs/2503.14456. arXiv:2503.14456.

18 Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. PaTH attention: Position encoding via accumulating Householder transformations, 2025. URL https://arxiv.org/abs/2505.16381.

19 Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning*, pages 21502–21521, 2024. URL https://proceedings.mlr.press/v235/jelassi24a.html.

20 Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Re. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1763–1840, 2024. URL https://proceedings.mlr.press/v235/arora24a.html.

3.4 Industry Applications and Multi-step Reasoning

Clayton Sanford (Google – New York, US)

This tutorial discussed connections (existing and potential) between applied LLM research and foundational theory of Transformers, with a particular focus on multi step reasoning, and the communication lens.

I first provided a list of guiding questions for empirically-influenced theory. These included questions related to *reasoning capabilities* (What are the "correct" reasoning primitives to learn and how can they be evaluated? How are these capabilities impacted by a shift from single-pass transformer inference to multiple prompts); *scaling laws* (How can theoretical results be framed as functions of compute budget, rather than context length? Can theoretical results incorporate power law assumptions?); and *model trade-offs* (What guidance can be provided for models operating under different resource constraints, e.g., on-device models whose train compute budget is less constrained than its parameter count, models that require regular training updates).

I then introduced multi-step reasoning (instantiated by the k-hop induction heads task) and provided a series of communication models that exploit capacity constraints in different modeling regime. *Two-party communication bounds* establish the hardness of expressing tasks like induction heads with single-layer transformers. The *blackboard communication protocol* is a multiparty communication model that establishes the limitations of subquadratic models (e.g. linear attention, kernel attention) to solve multi-step tasks. The *CONGEST* framework establishes the limitations of GNNs. The *Massively Parallel Computation (MPC)* model has a bidirectional relationship with deep transformers and provide provides (conditional) lower bounds on multi-step reasoning.

I concluded with two research vignettes.

- An explanation of a *star graph path detection* problem with hard observed learnability.
- A theoretical result that establishes separations for *mixture of experts (MoE)* models.

4 Working Groups

4.1 The Big Picture

Noah A. Smith (University of Washington – Seattle, US), George Cybenko (Dartmouth College Hanover, US), Ashish Sabharwal (Allen Institute for AI – Seattle, US), Gail Weiss (EPFL – Lausanne, CH)

This working group focused on the challenge of guiding theoretical researchers toward questions of relevance to those building modern language model-based systems. The group met for one day (Monday) before dispersing, and identified a few potential areas where more attention might lead to useful results. Some of these align well with longstanding ways of thinking about how theory can serve practice:

- Proofs of impossibility; identify capabilities that are unachievable, so that applied researchers do not waste time building systems to achieve them.
- Efficiency of learning: shedding light on design choices that are predicted to learn a capability most efficiently.

Other directions are more focused on specific practices that are being applied widely, or on desiderata that are being sought widely, in the language modeling community today:

- Theory of distillation (where a model is trained to simulate another model's behavior).
- Theory for interpretability, e.g., mapping a learned transformer into rules.
- Theory for capabilities of language models, e.g., developing abstractions for language model computation that are grounded in theory rather than analogy to human cognition, or theories that explain patterns of task interaction and interference during learning.
- Theory for evaluation that helps move evaluation practices toward better interpretability of model behaviors.
- Theory for training models (e.g., for scaling laws).
- Theory of large collections of data as they relate to learning and capabilities.

A general observation was that, in most scientific fields, "theory" refers to an evolving collection of propositions that explain phenomena, which is used to generate hypotheses that can be tested, and whose results then lead to updates to the theory. The consensus was that this loop is not yet well established; large-scale experimental work is driven more by the goal of improving benchmark performance than by advancing a theory of language models.

4.2 Learnability

Brian DuSell (ETH Zürich, CH), Gail Weiss (EPFL – Lausanne, CH), Michael Benedikt (University of Oxford, GB), George Cybenko (Dartmouth College Hanover, US), Robert Frank (Yale University, US), Anthony W. Lin (RPTU Kaiserslautern-Landau, DE), Paul S. Lintilhac (Dartmouth College Hanover, US), Jon Rawski (San José State University, US), Guillaume Rabusseau (University of Montreal, CA), Clayton Sanford (Google – New York, US), Noah A. Smith (University of Washington – Seattle, US), Lena Strobl (University of Umeå, SE), Andy Yang (University of Notre Dame, US)

This working group discussed theoretical and experimental approaches to understanding the learning dynamics of transformer networks. The group met from Monday through Friday, following an arc through theoretical to practically grounded discussions with varying participants.

Much theoretical work on neural networks has focused on their expressivity, i.e., the set of functions for which a parameterized network exists that implements them. Comparatively little work has formally characterized the *learnability* of functions by neural networks, i.e., the functions for which network network training algorithms have a reasonable probability of producing a network that implements them.

The discussion began with a recognition of the difficulty of capturing "learnability," especially as it applies to modern settings, e.g., the training of transformer neural networks with an adaptive, momentum-based optimizer (e.g., Adam). Rather than the more traditional setting of finding algorithms which could successfully learn a given task, we asked whether we could characterize the tasks learnable with a given algorithm. Simultaneously, we wondered whether it would be possible to characterize these tasks without getting completely entangled in the details of these complicated algorithms.

We discussed existing theoretical frameworks with which we could approach this problem, including in particular *sharpness* in the loss landscape of neural networks (the extent to which a minor perturbation in parameters will change the loss). We also discussed the expression of Boolean functions as linear combinations of sparse parities, properties of the class of *sensitive* languages (languages for which changing a single or small number of bits in an input is sufficient to change its classification) and *regular* languages (languages recognisable with a deterministic finite state automaton). We also discussed closure properties of learnability, for example the learnability of composed simple tasks.

On Thursday and Friday, we moved to exploring a case study involving two languages which individually are easy to learn, but very difficult when mixed together – despite the mixture permitting an apparently simple solution. The discussion was fruitful, raising multiple hypotheses and potential approaches to understanding the source of this added difficulty.

4.2.1 Discussed Problems

This group discussed problems related to loss landscapes (sharpness in parameter space), brittleness of transformer solutions, generalization, and interference. We detail our main discussions below.

4.2.1.1 Loss landscape of transformers computing regular languages

One broad question that was discussed is: how do different parameterizations of a DFA-simulating transformer affect the loss landscape around the solution?

More concretely, given a DFA, we can construct a transformer simulating it (up to some length n) using several techniques, for example:

- naive approach with n layers: each layer implements one transition of the DFA
- recursive scan approach with $\log n$ layers
- shortcut solution (if the DFA is solvable) with $\mathcal{O}(1)$ layers
- (if possible,) write a program recognizing the language in (some variant) of RASP and translate it into a transformer
- train a transformer on traces of the DFA

The first three approaches are described by [1].

Taking for granted that sharpness around the solution in parameter space is a good proxy for generalization ability / trainability, it is interesting to ask whether different parameterizations lead to different sharpness. Another dimension to consider is the structure of the DFA: how does its size, sparsity, number of components in the Krohn–Rhodes decomposition, etc., affect the sharpness of the different solutions. We hypothesized that in the case where a transformer permits multiple solutions to the same problem, the learning algorithm would prefer a solution with lower sharpness. However, we also considered that, due to the high complexity of the loss landscape, it may be possible that a specific data mix, task, or initialization could define a path into an otherwise sharp solution. We considered that sharpness may not tell the full story of learnability.

4.2.1.2 Interfering tasks when learning transformers

Gail shared an observation on two simple languages that are individually easy to learn, but together are much more difficult. These languages were the *copy* and *pairs* tasks, consisting of sequences of the sort ww for $w \in \Sigma^*$ (*copy*) and $w \in \Sigma_2^*$ for $\Sigma_2 = \{\sigma\sigma \mid \sigma \in \Sigma\}$ (*pairs*). We found that the union of these tasks (with a disambiguating prefix token) takes a long time

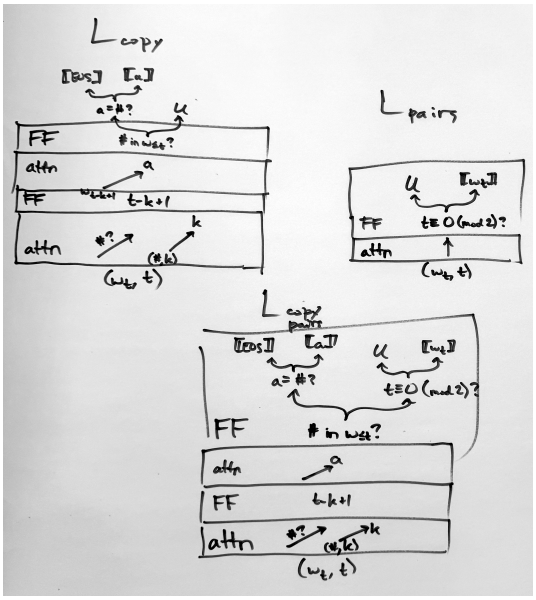


Figure 1 Simple transformers for copy (top left), pairs (top right), and copy-pairs (bottom). Sketch by Brian DuSell.

to learn jointly, much longer than the sum of their individual learning times. In particular, we observed that the pairs task is learned very quickly, while the copy task experiences significant delay (∼8x). However, when there is a warmup period where we learn only copy and then learn both, learning happens much faster, closer to what we would expect. We found that a composition of these languages called *copied pairs*, containing sequences of the sort $aabbcc :: aabbcc$, was also much slower to learn than its component tasks.

We found that both of these tasks are very easy to express in a transformer, and even described a simple solution for the copied pairs task in a simple 2 layer transformer. The trained transformers had 4 layers and 4 heads each, and generally sufficient representation space for the task, as evidenced by their ability to model it given enough time, as well as their ability to reach a solution quickly under helpful interventions. Hence the challenge must stem from some combination of the data presentation, learning algorithm, and model biases, rather than a limitation in expressive power. An interesting observation arose, in that once pairs was learned in these combined tasks (which would happen early on), it remained at near-zero loss for the duration of training, suggesting that from that point the learning algorithm was only able to explore the loss landscape where the pairs loss was low, restricting the exploration.

This is a concrete example of what we identified as a major open research question: **what kinds of formal languages have the property of being learnable under different kinds of compositions?** Said differently, which kinds of formal languages interfere with each other during training, either constructively or destructively? As a baseline, we can see that learning the copy task and reverse task together does not experience as much destructive interference. This question is completely separate from expressivity. But given some of the results on using formal languages in a learning curriculum to benefit natural language learning, it seems important to better understand these kinds of interference, and how it is affected by the order of training.

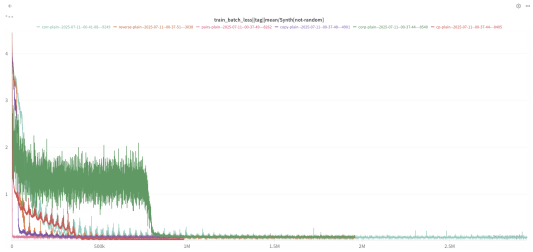


Figure 2 An image showing the loss over time (measured in number of training samples) while learning the combined copy-or-pairs task (green) as compared to just copy (purple) or just pairs (pink) tasks. The loss curve of the composed task, copied pairs, is shown in red. A brief exploration of reverse is also shown: reverse in orange, and copy-or-reverse in blue.

4.2.1.3 Other open problems & research questions

Other problems were discussed during the sessions of this group, including:

- Is the set of functions learnable by transformers closed under common operations (union, concatenation, etc.)?
- How sharpness can affect or be affected by decisions in training, in particular with how it can be involved in the training objective, and how it relates to batch size or other decisions.
- How architectural changes in a transformer can affect the learnability of different tasks, for example the use of rotary versus learned positional embeddings.

4.2.2 Possible Approaches

4.2.2.1 Loss landscape of transformers computing regular languages

For this question, two first steps were identified:

- Experiments. Collect a set of "interesting" automata (parity, flip-flop, bounded Dyck languages, etc.), translate each of them into a transformer using different constructions and estimate the sharpness of the loss landscape around the corresponding point in the parameter space. The sharpness can be estimated using a relatively naive Monte-Carlo approach (simply sample perturbed transfomers and compute the average loss difference over the train and test set). More efficient approaches could be investigated, such as random projections to estimate the trace of the Hessian.
- Theory. Deriving an analytical expression of the Hessian and its spectrum, which characterize the sharpness) is likely too difficult for even a small transformer construction with multiple layers. However, it would be interesting to formally analyze the sensitivity of the different building blocks used in different constructions. For example, in the shortcut solutions, each attention layer implements one of the simple automata from the Krohn–Rhodes decomposition while the MLP computes a form of parity function. Similarly, in the recursive scan solution, attention layers are used to map transition functions with one another while the MLP layers implement composition of these transitions. It should be possible to derive analytical results on the sensitivity of each of these building blocks to inform us on which one are more sensitive/brittle.

4.2.2.2 Case study on interference

Several interesting questions were raised to validate and explore the observation on copy-pairs interference, among them:

1. Training hyperparameters: to validate the observation, participants asked whether the interference holds even when when each task is hyperparameter-tuned? (i.e., could the delay be due simply to poorly adapted hyperparameters?)
2. Model Architecture: is the copy-pairs interference a special case of the architectures used in this observation, or does it hold for other sequence-processing models? In particular, does it hold both for transformers with learned and with rotary positional embeddings? And can a similar task pair be found for recurrent neural networks and their variants?
3. Task specifics: Was there something special about the composition of copy and pairs in the *copied pairsd*' task that encouraged the interference, which would not hold in compositions? For example, would pairing only on the left (sequences of the sort $aabbcc :: abc$) or right ($abc : aabbcc$) lead to similar levels of interference?
4. Model Generalisation: due to the nature of next-token based training, in which the model is only exposed to positive samples, a discussion arose around the exact rules the model is learning for its task. These can be explored via evaluating its response to out-of-domain inputs: for example, a well-trained *copied pairs* model, when presented with a malformed prefix such as $aabbcd ::$ (i.e., not adhering to left-pairing as expected), could reveal the model's tendency towards either pairs (generating *aabbcc*) or copy (generating *aabbcd*)?
5. Source of problem: the biggest question in the discussion was why this interference was happening at all, and what in the training was going wrong to delay copy learning. We raised several hypotheses, among them: degeneration of positional embeddings (as next-token predictions for the pairs task require only an understanding of the current token and current position's parity), or an initial strong devaluing of attention heads, due to their irrelevance for the pairs task? We considered comparing the gradients from different subtasks for these questions.
6. Characterisation: To gain a better understanding of the phenomenon, we raised the avenue of finding additional language pairs showing this interference, in particular with the hopes of finding enough to characterise and predict whether a pair of languages will or will not interfere in learning.

The group looks forward to exploring these questions in future work.

References

1 Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=De4FYqjFueZ`.

4.3 Interpretability

Michael Hahn (Universität des Saarlandes – Saarbrücken, DE), Anej Svete (ETH Zürich, CH), Joshua M. Ackerman (Dartmouth College – Hanover, US), Satwik Bhattamishra (University of Oxford, GB), Jiaoda Li (ETH Zürich, CH), Paul S. Lintilhac (Dartmouth College Hanover, US), Andy Yang (University of Notre Dame, US)

This working group focused on ways in which the theory of language models could support the development of interpretability methods of language models.

4.3.1 Discussed Problems

- Desiderata of an interpretation,
- How theoretical work can help interpretation efforts,
- Prospects for extracting logical descriptions from neural language models,
- Local vs. global interpretations,
- Formalization of what researchers usually refer to as "circuits for interpretation", and
- Compression of models for interpretability.

4.3.2 Possible Approaches

4.3.2.1 Desiderata of interpretability

A usable interpretation should satisfy multiple desiderata.

1. It should capture the right level of *abstraction* – it should allow human-readable information while staying faithful to the computation performed by the model,
2. It should be *actionable* – it should allow the user to intervene on the human-interpretable computations performed by the model and change them to induce target behavior,
3. It should enhance the *scientific understanding* of the model's computation, biases, etc.,
4. It should provide insights into the usability and limitations of the model – it should inform the user when the model's outputs can be trusted and when not,
5. It should enhance the user's trust in the model – its interpretation should provide evidence that the model is faithful to the user's intent, and
6. It should aid model's auditability and ensure that the model satisfies requirements for deployment.

4.3.2.2 Extracting the logical descriptions of neural language models

Existing literature provides exact characterizations of transformers as counter-free and partially-ordered finite-state automata. This theoretically allows us to convert any transformer into an automaton and leverage tools available for the interpretation of those on transformers. For instance, expressing a multiple transformers as finite-state automata would allow one to ask about their equivalence, differences, and emptiness. It would also allow one to test whether the models are provably performing computations required by looking for patterns in the automata.

Nevertheless, issues arise when considering the size of the resulting representations – the multi-step conversions required by known constructions result in super-exponential growth of the automata with respect to parameters such as the precision and number of layers. This is in conflict with the desideratum of a manageable and human-readable representation and raises questions about the feasibility of actual implementations.

4.3.2.3 Formalization of a circuit for interpretability

The group discussed the notion of *circuits* prevalent in recent interpretability research on language models. Such circuits provide *local* interpretations specific to individual input prompts x. Faithfulness to the original language model is typically evaluated on prompts formally highly similar to the target prompt x, often following some template. The group then discussed how one could formalize such local interpretations in the case of transformer language models, and desiderata for overall meaningfulness for such local interpretations.

4.3.2.4 Model compression for interpretability

Motivated by the desideratum of interpreting models with smaller but faithful representations, we discussed the problem of model compression. Intuitively, by compressing a model into a faithful smaller and thus easier-to-interpret representation, one could make inferences about the original model.

We discussed multiple avenues for compression:

- Quantization of model weights,
- Removing neurons in the feedforward networks, i.e., reducing the model width,
- Removing heads, which is analogous to reducing the model width,
- Low rank representations of the matrices, and
- Reducing the number of layers.

The last point seems challenging as it requires a large intervention on the model. Nevertheless, some results on the depth requirements to express certain languages do suggest that sometimes the depth could be reduced.

To address the other compression methods, the group discussed in what way the compressed model should be equivalent of similar to the original one. A good starting point could be requiring the compressed model to represent an equivalent *recognizer* – the identical string-to-membership function. Concretely, in the standard transformer setting in which real-valued outputs are mapped to classification decisions, one can take advantage of the classification gap to change the model slightly without influencing the string classifications.

4.4 Uniformity

William Merrill (New York University, US), Laura Strieker (Leibniz Universität Hannover, DE), Satwik Bhattamishra (University of Oxford, GB), Michaël Cadilhac (DePaul University – Chicago, US), David Chiang (University of Notre Dame, US), Ashish Sabharwal (Allen Institute for AI – Seattle, US), Clayton Sanford (Google – New York, US), Howard Straubing (Boston College, US)

Much work on transformer expressivity has followed in the formal language theory tradition of defining a problem as expressible if there exists a fixed transformer (with respect to the input length n) that can recognize a language. However, some work allows the hyperparameters of the network (e.g. depth, width) or even the parameter values themselves to depend on n. To understand the differences between these regimes, this group sought to carefully define a notion of *uniformity* for transformers whose parameters can depend on the sequence length. We used this definition to instantiate several natural classes of uniform transformer families

generalizing the standard fully uniform transformers that do not depend on n in any way. We also considered potential separations between these classes (e.g., between fully and L-uniform classes).

4.4.1 Discussed Problems

We propose carefully defining and analyzing the uniformity of families of transformers, motivated by the following questions:

- How should we think about uniformity for transformers? What differences are there to circuit classes.
- Theories of expressivity with a bounded sequence length
- Better accounting for the role of depth and width in transformer expressivity
- Separations between transformers that are fully uniform and transformers where some parameters or hyperparameters can depend uniformly n

4.4.2 Possible Approaches

We can think about two ways that a transformer can potentially change as the sequence length n increases:

1. The parameters θ themselves can change. The shape could remain the same but values can adapt. Potentially, the number of parameters could even expand with n.
2. Given a fixed vector of parameters, aspects of inference (how we apply the model to data) could change. For example, we could increase the precision as $O(\log n)$ or pad the input with $O(\text{poly}(n))$ blank tokens. Crucially, in either case, we use the same parameters for all input lengths.

To account for both of these forms of change with n, we propose decomposing a transformer into its architectural schema and parameter vector. The architectural schema can specify that certain properties of inference evolve with n (e.g., precision, chain of thought, or padding). Uniformity of the transformer weights can then be defined similarly to circuits.

Concretely, we can define an architectural schema as follows:

▶ **Definition 1.** A *transformer architecture* $\mathcal{T}$ is a Tuple $(p, t, d, w, e, m, n, a, c)$ where

- p denotes the precision,
- t temperature,
- d depth,
- w width,
- e positional encoding,
- m masking,
- n layer norm,
- $a \in \{uha, aha, sma\}$ attention,
- c number of chain of thought steps (or padding tokens).

We call these the *hyperparameters* of a transformer.

The following example illustrates this definition. $\mathcal{T} = (O(1), ..., O(1), sma, 0)$ are fixed-precision, temperature, and width transformers with softmax attention and no chain of thought. None of the hyperparameters depend on n.

A transformer architecture $\mathcal{T}$ can be parameterized by a weight vector θ to obtain a language recognizer $\mathcal{T}_\theta : \Sigma^* \to \{0, 1\}$. We can then define uniform transformer classes in terms of these language recognizers and uniform families of parameters $\{\theta_N\}_{N=0}^{\infty}$:

▶ **Definition 2.** Let $\mathcal{T}$ be a transformer architecture. Then U-uniform-$\mathcal{T}$ is the set of languages L such that there exists a family of parameter vectors $\{\theta_N\}_{N=0}^{\infty}$ such that $\mathcal{T}_{\theta_N}$ recognizes $L^{\leq N} = \{w \mid |w| \leq N \wedge w \in L\}$ for all N and the function $N \mapsto \theta_N$ can be computed in the class U.

Note it is enforced that, as we change the parameters of a transformer family, its behaviour must remain the same as smaller inputs. This restriction is not possible with circuits, which take a fixed-size input. In contrast, transformers can read a variable-length string as input.

Let $\mathcal{C}$ denote the set of constant functions $\mathbb{N} \to \Sigma^*$ and recall $\mathcal{T}_{fu}$. Then $\mathcal{C}$-uniform-$\mathcal{T}_{fu}$ is the expressive power of "fully uniform" transformers where the parameters cannot change at all with n. In contrast, $\mathcal{L}$-uniform-$\mathcal{T}_{fu}$ is the class where the number of parameters stays the same, but the values of θ can change uniformly in a way computable in log space. Finally, let $\mathcal{T}_w = (O(1), ..., O(1), O(\log n), sma, 0)$ denote transformers with logarithmic width. Then $\mathcal{L}$-uniform-$\mathcal{T}_w$ is the set of languages recognizable by transformers with logarithmic width, where the parameters can change with n in a way computable in log space. Thus, our single definition can be instantiated to make all of these notions of uniform transformer families precise.

4.4.3 Tentative Separations

A general theme is that, if a fully uniform transformer model $\mathcal{C}$-uniform $\mathcal{T}$ has $O(\log n)$ communication complexity, we expect that L-uniform $\mathcal{T}$ will not. This can allow us to show a separation between the $\mathcal{C}$-uniform and L-uniform $\mathcal{T}$.

- One layer: communication complexity lower bound for fully uniform version; increase temperature to increase expressive power for L-uniform model
- Multilayer: [3] gives analogous communication complexity lower bound; increase temperature to gain power for L-uniform model
- C-RASP: same idea of $O(\log n)$ communication complexity, but uniformity lets us expand that
- What about C-RASP transformers with dynamic temperature? Does the additional power of L-uniformity persist?
- With padding, L-uniform and fully uniform AHAT collapse, since we get both equal to TC^0.

4.4.4 Conclusions

We have made progress towards a general and rigorous notion of uniformity for families of transformers, which we hope will aid future work studying the expressivity of transformer families, allowing it to be more standardized and rigorous. An immediate next step is refining our understanding of separations between various types of transformers with different levels of uniformity, which could have some relevant for understanding fundamental limits on length generalization: inexpressibility of a language under a fully uniform model implies length generalization is not possible on that language. More generally, better understanding the expressivity of uniform transformers can help us better conceptualize expressivity with a maximum context length and the role of parameters like depth or width in transformer expressivity.

4.5 Depth

Satwik Bhattamishra (University of Oxford, GB), Clayton Sanford (Google – New York, US), Michael Hahn (Universität des Saarlandes – Saarbrücken, DE), Ashish Sabharwal (Allen Institute for AI – Seattle, US), Andy Yang (University of Notre Dame, US)

Understanding the limitations of small-width multi-layer Transformers has remained challenging. For sequences of length $\leq N$, recent works [1, 2] show that one-layer Transformers of width $o(N)$ cannot express certain functions, but analogous limitations for multi-layer models are still elusive. An interesting open direction is to identify functions that can be represented by three-layer Transformers of small width (e.g., $O(\log N)$) but not by two-layer Transformers (e.g., they require $\omega(\log N)$ width). Such separations are known for 1-layer vs. 2-layer Transformers [2]. Notably, [3] proved unconditional lower bounds and provided separations for multi-layer Transformers with *causal masking*; comparable lower bounds are not known for Transformers without causal masking.

4.5.1 Discussed Problems

- Our primary focus was on approaches to derive unconditional lower bounds for multi-layer Transformers.
- We explored potential strategies beyond the current communication complexity–based reductions.
- We compared and contrasted the logic-based and communication complexity-based results on transformers.
- We sought to identify concrete functions or tasks that appear easier for three-layer Transformers but not for two-layer Transformers as a starting point for such analysis.
- We discussed barriers that make proving lower bounds for two-layer Transformers harder than for one-layer models.

4.5.2 Open Problems and Possible Approaches

KW games. One approach we explored is adapting Karchmer–Wigderson (KW) games [4], a communication-complexity technique that has been fruitful for deriving depth lower bounds for monotone Boolean functions. Current approaches partition the input between two parties and show that the network output can be computed with a few bits of communication (depending on the network width), which then yields width lower bounds. In contrast, KW games organize the protocol so that communication proceeds *across depth*, and CC-reductions lead to depth lower bounds. When attempting to apply this idea to Transformers, we encountered barriers: for circuits, the approach relies (to some degree) on monotonicity (which does not hold for neural nets) and on bounded fan-in. The latter holds for hard-attention Transformers but not more general settings. Developing a nontrivial adaptation that leverages the KW-games paradigm [4] for Transformers remains an interesting open direction.

2 vs. 3 layer separation. As a starting point, we aimed to identify tasks that are intuitively difficult for two-layer Transformers but not for three layers. The goal is to isolate a concrete problem that can be analyzed to formalize these intuitions.

We develop the *compositional indexing* problem (Def. 1), based on the following idea. A Transformer layer consists of two blocks: the attention block, which gathers or processes information across the sequence, and the feedforward block, which processes information at a particular position. Prior work shows that certain tasks (e.g., Inner Product mod 2, IP2, and Disjointness) are hard for one-layer Transformers but not for two layers. We extend this by asking the model to compute IP2 not on the raw input $x \in \{0,1\}^N$ but on a subsequence specified by a list of indices $i_1, \ldots, i_N \in [N]$. The model receives a bit string $x_1, \ldots, x_{2N} \in \{0,1\}$ followed by indices $i_1, \ldots, i_N \in [2N]$, and must compute IP2 on the selected bits. The formal definition is as follows.

▶ **Definition 1** (Compositional indexing problem)**.** Fix a positive integer $N \in \mathbb{N}$ and write $[m] = \{1, 2, \ldots, m\}$.

Base function. Let $f\colon \{0,1\}^N \to \{0,1\}$ be any Boolean function with two-party communication complexity $\Omega(N)$. For concreteness, one may take f to be the IP mod 2 function or Equality. The IP mod 2 function $\mathrm{IP2} : \{0,1\}^{N/2} \times \{0,1\}^{N/2} \to \{0,1\}$ is

$$\mathrm{IP2}(\mathbf{x}, \mathbf{y}) \;=\; \langle \mathbf{x}, \mathbf{y} \rangle \bmod 2 \;=\; \big(h(x_1, y_1) + h(x_2, y_2) + \cdots + h(x_{N/2}, y_{N/2})\big) \bmod 2,$$

where $h(x_i, y_i) = x_i y_i$.

Selection operator. Define

$$\mathrm{Sel}_N\colon \{0,1\}^{2N} \times [2N]^N \;\longrightarrow\; \{0,1\}^N, \qquad \big(\mathrm{Sel}_N(x, \mathbf{i})\big)_j \;=\; x_{i_j} \quad \text{for } j = 1, \ldots, N,$$

where $x = (x_1, \ldots, x_{2N}) \in \{0,1\}^{2N}$ and $\mathbf{i} = (i_1, \ldots, i_N) \in [2N]^N$. Thus, Sel_N outputs the length-N subsequence of x at coordinates $i_1, \ldots, i_N$.

Composed function. The final mapping is the composition

$$g\colon \{0,1\}^{2N} \times [2N]^N \;\longrightarrow\; \{0,1\}, \qquad g(x, \mathbf{i}) \;=\; f\big(\mathrm{Sel}_N(x, \mathbf{i})\big).$$

In other words, g receives a length-$2N$ binary string together with N indices, extracts the indexed bits via Sel_N, and feeds the resulting length-N string into f. Let CoIP2 denote the compositional IP mod 2 function,

$$\mathrm{CoIP2}(x, \mathbf{i}) \;=\; \mathrm{IP2}\big(\mathrm{Sel}_N(x, \mathbf{i})\big).$$

For the compositional indexing problem, it is straightforward to adapt techniques from [2] to show that three-layer Transformers with width $O(\log N)$ can compute CoIP2. Similarly, two-layer Transformers with width $O(N)$ can compute it. The key open question is whether two-layer Transformers with width $o(N)$—or even $O(\log N)$—can compute CoIP2. Proving that two-layer Transformers require width $\omega(\log N)$ would yield a separation between two and three layers.

In this problem, the model receives the pair $(x, \mathbf{i})$. Upon seeing the indices $\mathbf{i} = (i_1, \ldots, i_N)$, if the first layer only retrieves the input bits $x_{i_1}, \ldots, x_{i_N}$, then we have that the second layer cannot compute functions like IP2 based on prior works. However, the first layer need not be restricted to this behavior – this is the main challenge. A useful observation is that to compute functions like IP2, the attention block must first gather input bits at paired indices $(i_1, i_{\frac{N}{2}+1}), \ldots, (i_{\frac{N}{2}}, i_N)$. When the Transformer receives the input index i_k, while it can retrieve the bit x_{i_k}, one can show (by adapting arguments from [5]) that it cannot simultaneously retrieve bits x_{i_j} for any other index i_j with $j \neq k$. Thus one can prove that the first layer cannot perform the one-hop retrieval or index lookup, which could be potentially useful for proving lower bounds for the second layer. For the reasons described above, we hypothesize that CoIP2 could be hard for log-width two-layer Transformers but the problem remains open.

References

1 Clayton Sanford, Daniel J Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems*, 36:36677–36707, 2023.
2 Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
3 Lijie Chen, Binghui Peng, and Hongxun Wu. Theoretical limitations of multi-layer transformer. *arXiv preprint arXiv:2412.02975*, 2024.
4 Mauricio Karchmer and Avi Wigderson. Monotone circuits for connectivity require super-logarithmic depth. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 539–550, 1988.
5 Clayton Sanford, Daniel Hsu, and Matus Telgarsky. One-layer transformers fail to solve the induction heads task. *arXiv preprint arXiv:2408.14332*, 2024.

4.6 Probability

David Chiang (University of Notre Dame, US), Ryan Cotterell (ETH Zürich, CH), Jiaoda Li (ETH Zürich, CH), Anthony W. Lin (RPTU Kaiserslautern-Landau, DE), Jon Rawski (San José State University, US), Noah A. Smith (University of Washington – Seattle, US), Andy Yang (University of Notre Dame, US), Anej Svete (ETH Zürich, CH)

4.6.1 Discussed Problems

Most theoretical papers on transformers treat them as language recognizers, where the input is a string and the output is a truth value (true for accept, false for reject). But transformers are usually used as *language models*, which define a probability distribution $P(w_t \mid w_1 \cdots w_{t-1})$ where $w_t \in \Sigma \cup \{\text{EOS}\}$. Instead of asking what languages are recognized by transformers, should we be asking what probability distributions are modeled by transformers?

We dug into this question, focusing on UHATs, which, as language recognizers, were shown by [1] to be equivalent to B-RASP and LTL. (Other transformer variants were also discussed; C-RASP in particular led to a spin-off topic in Section 4.9.)

4.6.2 Possible Approaches

We began by defining a generalization of B-RASP which maintains its restrictions while adding probabilities (or, more generally, weights), which we call PB-RASP. We examined various ways of relating them to probabilistic automata, and therefore to distributions over strings.

4.6.3 Conclusions

The definition of PB-RASP that we arrived at has much in common with the weighted first-order logic (FO) of [2]. Both have three layers: first, an unweighted logic (B-RASP or FO); second, the ability to choose weights conditioned on unweighted formulas, defining "step" functions; third, multiplication of weights across all positions. One advantage of this formulation is that it cleanly separates two differences between language recognizers and

language models: on the one hand, language models are *weighted* while language recognizers are unweighted, and on the other hand, language models are *autoregressive* while language recognizers are not.

As language recognizers, UHATs and B-RASP are equivalent to counter-free automata, but as language models, counter-free automata split into (at least) two levels of expressivity, counter-free DFAs and counter-free NFAs. A key finding of this working group is that PB-RASP, and therefore language models based on UHATs, are equivalent to counter-free DFAs, not NFAs.

Follow-up discussions of this question have led to further developments, which we hope to write about elsewhere in the very near future.

References

1 Andy Yang, David Chiang, and Dana Angluin. Masked hard-attention transformers recognize exactly the star-free languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 10202–10235, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/13d7f172259b11b230cc5da8768abc5f-Paper-Conference.pdf`.

2 Manfred Droste and Paul Gastin. Aperiodic weighted automata and weighted first-order logic. In *Proceedings of the 44th International Symposium on Mathematical Foundations of Computer Science*, 2019. doi: `10.4230/LIPICS.MFCS.2019.76`.

4.7 Recurrence

Gail Weiss (EPFL – Lausanne, CH), Brian DuSell (ETH Zürich, CH), Robert Frank (Yale University, US), Martin Grohe (RWTH Aachen, DE), Laura Strieker (Leibniz Universität Hannover, DE)

This working group discussed recurrence as it relates to the expressive and practical power of neural language models. The group met for two sessions on Tuesday, identifying some potential future experiments before dispersing.

4.7.1 Discussed Problems

Our discussion distinguished between two types of recurrence: *horizontal*, in which a model updates a state for every input token in an ordered sequence, and *vertical*, in which the model may continue to increase its computation depth in response to its own current state. Martin shared insight on the use of vertical recurrence for evaluable problems such as minimal graph colouring, while Brian shared results on the direct advantages of horizontally recurrent models over attention-based non-recurrent ones [1, e.g.]. We discussed how the advantages (and weaknesses) of recurrence could impact the processing of natural language sequences specifically. We discussed the natural language motivations for horizontal recurrence, and how we may approach these motivations while maintaining the powerful parallelisation abilities of the non-recurrent transformer. The discussion also touched on potential limitations of the causal attention mask used in autoregressive transformers.

- What problems are vertical and horizontal recurrence used to solve, and are these to be expected in natural language?

- What is the performance and efficiency of modern recurrent or semi-stateful models, and how successful are current attempts at integrating statefulness into modern models?
- What are the main weaknesses of non-recurrent models as they relate to natural language? What are the main weaknesses of recurrent models?
- What methods other than recurrence are there to improve length generalization in transformers?
- What other capabilities of language models have been sacrificed in the name of parallelisation?

A related interesting point that arose during the discussion was the impact of the causal attention mask deployed in transformer-based natural language modeling, which allows high parallelization during training at the cost of further reduced processing capacity over individual input tokens. In this setting, the embeddings for early tokens in a sequence are based on far less of the input than they could be, even when being used for predictions that may read the entire sequence.

4.7.2 Possible Approaches

Reflecting on the combined issues of lacking recurrence and limiting causal mask, we realised that the processing of a long input sequence could potentially be improved by applying horizontally recurrent computation over the sequence, with varying computation depth for different tokens and a gradually lifted causal mask, allowing the processing of earlier sub-blocks to shift from local next-token predictions to fuller-sequence embeddings. We left the careful implementation and exploration of such an architecture to future work.

Other motivations that arose for recurrence involved length generalization of networks to long inputs, and the ability to perform arbitrarily deep computations (depth generalization). For these, we found several works that may inspire future experiments, covering for example: the incorporation of horizontal (Jamba, [2]) and vertical (universal transformers, [3]) recurrence, or similar temporal bias (Transformer-XL, [4]), into transformers, works on stateful but highly parallelisable models (e.g., S4, [5]), and even interesting works on parallelizing the implementation of recurrent neural networks (RNNs) while not impacting their actual recurrence, to allow their more effective scaling [6].

4.7.3 Conclusions

The linearly increasing computation depth of recurrent models provides them clear advantages in certain tasks, but the training routines for such models are not easily parallelisable, and hence they are difficult to exploit with today's resources. Nevertheless, a pragmatic targeting of the problems we wish to improve on for natural language processing specifically may allow for more feasible solutions.

References

1 William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjNGlPh8Wh.

2 Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, et al. Jamba: Hybrid Transformer-Mamba language models. In *Proceedings of ICLR*, 2025. URL https://openreview.net/forum?id=JFPaD7lpBD.

3 Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *Proceedings of ICLR*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

4 Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285/.

5 Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *Proceedings of ICLR*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

6 Yi Heng Lim, Qi Zhu, Joshua Selfridge, and Muhammad Firmansyah Kasim. Parallelizing non-linear sequential models over the sequence length. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=E34AlVLN0v.

4.8 Chain of Thought

Ashish Sabharwal (Allen Institute for AI – Seattle, US), William Merrill (New York University, US), Michaël Cadilhac (DePaul University – Chicago, US), Howard Straubing (Boston College, US), Laura Strieker (Leibniz Universität Hannover, DE), Michael Hahn (Universität des Saarlandes – Saarbrücken, DE)

This working group focused on how the amount of intermediate generation (i.e., the number of Chain of Thought tokens) affects the representation power of Transformers, especially in the sublinear regime. It is known that, on inputs of size n, $\mathsf{CoT}[\mathrm{poly}(n)] = \mathsf{P}$ and $\mathsf{CoT}[n]$ can solve certain NC^1-complete problems. On the other hand, allowing only $O(\log n)$ intermediate steps doesn't seem to add more expressive power: $\mathsf{CoT}[\log n] \subseteq \mathsf{TC}^0$, which coincides with the best known upper bound on $\mathsf{CoT}[0]$, i.e., no intermediate generation at all. What interesting things can we say about $\mathsf{CoT}[f(n)]$ for sub-linear functions f?

4.8.1 Discussed Problems

- Can $\mathsf{CoT}[\sqrt{n}]$, or, more generally, $\mathsf{CoT}[n^{1/c}]$, solve problems beyond TC^0?
- Can $\mathsf{CoT}[\log^2 n]$, or, more generally, $\mathsf{CoT}[\log^k n]$, solve problems beyond TC^0?

Consider a "parallel" version of intermediate generation, denoted $\mathsf{CoT}[g(n), f(n)]$, where the model takes $f(n)$ steps of generation and in each step gets to write $g(n)$ tokens (in parallel) rather than a single token, which is related to the practical method of speculative decoding [1].

- What can $\mathsf{CoT}[\mathrm{poly}(n), \log^k n]$ solve?

4.8.2 Possible Approaches

Going from linear to $n^{1/c}$ is possible via the standard *padding construction* from complexity theory. Let Q be an NC^1-complete language that is also in $\mathsf{CoT}[n]$, i.e., can be recognized using n CoT steps. Consider the language $Q' = \{w\, 1^{|w|^c} \mid w \in Q\}$ where 1 is some symbol

not used in Q, i.e., every string in Q is appended with polynomially many 1's. Then (a) Q' is also NC^1-complete and (b) can be recognized by $\mathsf{CoT}[n^{1/c}]$.

In fact, the same applies even when L is a P-complete problem. It follows that:

- For any $c > 0$, there is a P-complete problem in $\mathsf{CoT}[n^{1/c}]$.
- If $\mathsf{NC} \neq \mathsf{P}$, then there exists a language L that, for any $c > 0$ and $k \in \mathbb{Z}_{\geq 0}$, is in $\mathsf{CoT}[n^{1/c}]$ but not in $\mathsf{CoT}[\log n]$.

This should also extend in some form to the parallel version of CoT, i.e., $\mathsf{CoT}[\mathrm{poly}(n), n^{1/c}]$ vs. $\mathsf{CoT}[\mathrm{poly}(n), \log^k n]$.

It should be possible to show:

- $\mathsf{CoT}[\mathrm{poly}(n), \log^k n] \subseteq \mathsf{TC}^k$ by viewing a length-d CoT as a circuit of depth d.
- $\mathsf{CoT}[\mathrm{poly}(n), \log^k n] \supseteq \mathsf{NC}^k \supseteq \mathsf{TC}^{k-1}$, via NC^k's equivalence to alternating log-space, $\log^k$-time Turing machines.

For poly-log CoT, we have the following:

- Adding n many $(\log n)$-bit numbers is in $\mathsf{CoT}[\log^2 n]$. The idea is to first compress this addition to adding $\log n$ $(\log n)$-bit numbers, then add the resulting $\log n$ numbers sequentially. This problem is clearly in TC^0 but it's not known whether fixed depth transformers can solve it.

4.8.3 Conclusions

We considered the expressivity landscape for transformers with sublinear CoT, identifying critical regimes and open questions, and answering some of them. In particular, we showed that $\mathsf{CoT}[\sqrt{n}]$, or more generally, $\mathsf{CoT}[n^{-c}]$, can solve P-complete problems. Below this, we know that $\mathsf{CoT}[\log n] \subseteq \mathsf{TC}^0$. The intermediate regime of $\mathsf{CoT}[\log^k n]$ remains an interesting open question, as well as more rigorously analyzing CoT steps interwoven with parallel generation steps.

References

1 Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19274–19286, 2023. URL `https://proceedings.mlr.press/v202/leviathan23a.html`.

4.9 Automata

Andy Yang (University of Notre Dame, US), Michaël Cadilhac (DePaul University – Chicago, US), Ryan Cotterell (ETH Zürich, CH), Michael Hahn (Universität des Saarlandes – Saarbrücken, DE), Anthony W. Lin (RPTU Kaiserslautern-Landau, DE)

This working group focused on developing an automata-theoretic characterization of C-RASP and associated questions. This question started as an off-shoot of the probability group, looking ahead towards creating a probabilistic C-RASP.

First, there was discussion of characterizing C-RASP using counter/Parikh automata. A potential way forward would be to restrict to the class of partially ordered automata while adding some counting abilities. Next, a connection to typed wreath products of $(\mathbb{Z}, \mathbb{Z}_+)$ was also discussed. However, no conclusive results were obtained.

An effort to characterize the regular languages in C-RASP developed out of this discussion. It is known that C-RASP contains all R-trivial languages and all bounded-Dyck languages. It is also known that C-RASP avoids $\Sigma^* bb \Sigma^*$ and all non-aperiodic languages. An exact characterization remains elusive.

5 Other Questions

Below is a list of all the topics that were proposed for working groups. Many were merged to form the working group topics. Others were not discussed but would be valuable for future discussion. Some of these topics have been paraphrased.

- Definitions
  - How can we get some intuition into these architectural variations? (Michael Benedikt)
  - Better transformer definitions (Andy Yang?)
- The Big Picture
  - How do we convince engineering people that theory is important? Motivating examples, etc. (Brian DuSell)
  - Setting up for a good week: surveying what's there and seeing what's missing in expressivity (Laura Strieker)
  - What theory would be most important/relevant/useful to practical implementations of NLMs? (George Cybenko)
  - Give me something I can use to make strong models, cheaper, either from scratch, or after training. (Noah A. Smith)
  - What makes a successful "Theory of Neural Networks"? Is our work approaching that? (Andy Yang)
  - Identify (and start to fill) *gaps* between theoretical models we analyze and actual models practitioners deploy. (Ashish Sabharwal)
  - What transformer questions will take 100 years to solve? (Andy Yang)
- Learnability
  - What are the inductive biases of neural architectures (transformers, RNNs, etc.) when we ignore length generalization? (E.g., recognition up to length N.) How do we measure this? E.g., sample complexity. (Brian DuSell)
  - Learning vs. generalizing vs. expressing (Michael Hahn)
  - What is a canonical form of a transformer class to get a learning result? For example, a canonical DFA, C-RASP, etc. (Jon Rawski)
  - "Learnable Languages" but more rigorous. (Jon Rawski)
  - Interference: explaining difficulty learning two tasks, simultaneously and understanding which task pairs might interfere
  - Learning linear combinations of sparse automata. Expressibility and learnability. (Paul S. Lintilhac)
  - Can we systematically map natural language tasks to formal language complexity properties (degree, sensitivity, circuit depth) to predict generalization? (Paul S. Lintilhac)
- Interpretability
  - How can theory help interpretability? (Michael Hahn)
  - Formalizations, theory, practice, connection to formal verification, complexity theory, automata learning (Anthony W. Lin)
  - Can we translate Transformers to logic and use this for interpretability and/or formal verification? (William Merrill)

- Verification of transformers, e.g., how hard is it to test whether a given UHAT is equivalent to a given DFA? (David Chiang)
- Can we prove/make more rigorous these informal statements about limitations with respect to compositions of functions with formal languages? (Paul S. Lintilhac)
- Interpretability. How can we use theoretical results, constructions, and derivations for (coarsely) interpreting trained models? (Anej Svete)
- What's an "interpretation" I would like or find useful or insightful? (Gail Weiss)

- Continuity and probability
  - Transformers on time series? (Anthony W. Lin)
  - Rational or real numbers. Are SSMs with arbitrary precisions still in uniform TC^0? What about other networks? (David Chiang)
  - Language models as weighted/probabilistic logics, automata, etc. (Jon Rawski)
- Recurrence
  - Can we parallelise training for computation depth? (Gail Weiss)
  - Recurrence bias transformers still parallel? (Gail Weiss)
- Uniformity
  - Why are some results relying on nonuniform complexity classes? What are the uniform/nonuniform variants of the existing nonuniform/uniform results? (Michaël Cadilhac)
  - How different axes of increasing memory influence expressivity (Satwik Bhattamishra)
  - Constant-context-length transformer complexity: what is the right scale parameter if not context length? (Clayton Sanford)
  - Motivating, clustering, and relating definitions. e.g., what concerns motivate "full uniformity"? Philosophical? Practical? (William Merrill)
  - What languages are expressible up to a maximum length but not all lengths? (David Chiang)
  - How do we account for complexity parameters beyond n? Number of parameters, training time, number of experts, etc. Related to uniformity? (William Merrill)
- Depth
  - Depth? Unconditional separations without masking? Relationship to chain of thought? (Clayton Sanford)
  - Can we develop other techniques to derive lower bounds for transformers: multilayer or hybrid architectures? For example, static encodings, chronograms, counting linear regions (Satwik Bhattamishra)
- Other questions
  - Moment computation over attention (The "real" soft attention): Many of our constructions use hard attention or close simulations, but a lot of attention units look like partitions/clusters. Is there a theory that includes this? (Clayton Sanford?)
  - A lot of attention has been paid to transformer variants that have better time complexity. What about variants that have *worse* time complexity? (Gail Weiss)
  - Are residual connections necessary for the full expressiveness of transformers? (Pablo Barcelo)
  - Can we view decoders as strictly local transducers? (Jon Rawski)
  - Why are transformers so good at natural language but so bad at formal languages? (Brian DuSell)
  - Big model $\rightarrow$ Small model
  - Does monotonicity help? (Michaël Cadilhac)
  - Can we overcome limitations by configuring the network to the task? (Paul S. Lintilhac)

Participants

- Joshua M. Ackerman
Dartmouth College – Hanover, US
- Pablo Barcelo
PUC – Santiago de Chile, CL
- Michael Benedikt
University of Oxford, GB
- Satwik Bhattamishra
University of Oxford, GB
- Michaël Cadilhac
DePaul University – Chicago, US
- David Chiang
University of Notre Dame, US
- Ryan Cotterell
ETH Zürich, CH
- George Cybenko
Dartmouth College Hanover, US
- Brian DuSell
ETH Zürich, CH
- Robert Frank
Yale University, US
- Martin Grohe
RWTH Aachen, DE
- Michael Hahn
Universität des Saarlandes – Saarbrücken, DE
- Jiaoda Li
ETH Zürich, CH
- Anthony W. Lin
RPTU Kaiserslautern-Landau, DE
- Paul S. Lintilhac
Dartmouth College Hanover, US
- William Merrill
New York University, US
- Guillaume Rabusseau
University of Montreal, CA
- Jon Rawski
San José State University, US
- Ashish Sabharwal
Allen Institute for AI – Seattle, US
- Clayton Sanford
Google – New York, US
- Noah A. Smith
University of Washington – Seattle, US
- Howard Straubing
Boston College, US
- Laura Strieker
Leibniz Universität Hannover, DE
- Lena Strobl
University of Umeå, SE
- Anej Svete
ETH Zürich, CH
- Gail Weiss
EPFL – Lausanne, CH
- Andy Yang
University of Notre Dame, US

Report from Dagstuhl Seminar 25291

(Actual) Neurosymbolic AI: Combining Deep Learning and Knowledge Graphs

Pascal Hitzler*[1], Cogan Shimizu*[2], Daria Stepanova*[3], and Frank van Harmelen*[4]

1 **Kansas State University – Manhattan, US.** phitzler@googlemail.com
2 **Wright State University – Dayton, US.** cogan.shimizu@wright.edu
3 **Bosch Center for AI – Renningen, DE.** daria.stepanova@de.bosch.com
4 **VU Amsterdam, NL.** frank.van.harmelen@vu.nl

Abstract

In the past decade, both deep learning (DL) and knowledge graphs (KGs) have seen astonishing growth and groundbreaking milestones – DL due to newly available resources (e.g., accessibility of (modern) web scale data), previously un-scalable techniques (e.g., transformers), and modern hardware; KGs due to successful standardization, web-scale integration, and previously un-scalable techniques for querying and inference. This has brought new and increased interest to both fields, and especially in how they can complement each other. This report documents the program and the outcomes of Dagstuhl Seminar 25291 "(Actual) Neurosymbolic AI: Combining Deep Learning and Knowledge Graphs". This Dagstuhl Seminar brought 34 internationally recognized experts together to examine the gap between deep learning and knowledge graphs, and architect their integration: neurosymbolic AI.

Seminar July 13–18, 2025 – http://www.dagstuhl.de/25291
2012 ACM Subject Classification Information systems → Semantic web description languages; Theory of computation → Automated reasoning; Theory of computation → Description logics; Theory of computation → Semantics and reasoning; Human-centered computing → Human computer interaction (HCI); Computing methodologies → Machine learning
Keywords and phrases deep learning, knowledge graphs, neurosymbolic ai
Digital Object Identifier 10.4230/DagRep.15.7.53

1 Executive Summary

Cogan Shimizu (Wright State University, US, cogan.shimizu@wright.edu)
Pascal Hitzler (Kansas State University, US, phtizler@googlemail.com)
Daria Stepanova (Bosch Center for AI, DE, daria.stepanova@de.bosch.com)
Frank van Harmelen (VU Amsterdam, NL, frank.van.harmelen@vu.nl)

Run un-conference style, with merely three set presentations for topic introductions, the participants decided on themes of discussion groups within the theme of the seminar, and on the goal of providing a written account, found in this report, of the emerging themes, structured into definitions, ambitions, challenges, and the state of the art. The themes that emerged are the themes of the Breakout Group Reports found herein: Defining Neurosymbolic Systems; Symbol Emergence; Small Data and Neurosymbolic AI; Explainable AI; Neurosymbolic AI in the Age of Generative AI; Knowledge Graphs and Ontologies in Neurosymbolic

* Editor / Organizer

(Actual) Neurosymbolic AI: Combining Deep Learning and Knowledge Graphs, *Dagstuhl Reports*, Vol. 15, Issue 7, pp. 53–123
Editors: Pascal Hitzler, Cogan Shimizu, Daria Stepanova, and Frank van Harmelen
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Figure 1 A picture of the NeSy Theory cluster of topics. The specific colors do not have meaning. The dots signify a "vote" rather than repeating a new version of the sticky.

Systems; Cognition and Neurosymbolic AI; Benchmarks in the Neurosymbolic Ecosystem; and Real-World Applications in Neurosymbolic Artificial Intelligence. Additional discussions evolved around the general question of how the two major outlets for Neurosymbolic AI – the Neurosymbolic Learning and Reasoning Conference,[1] and the Neurosymbolic AI journal[2] – can best support the nascent community. Key drivers of both outlets were in attendance at the seminar and have already begun to set some of the discussion results into motion.

The Seminar

Thirty-four participants were finally able to attend the seminar. We believe that the diversity of our attendees was both fair, broad, and representative of the breadth of the fields: bridging seniority, gender diversity, geographic location, industry vs. academia, and expertise in neural or symbolic (or both) AI systems. But even through the variety, there were interesting through-lines and other connections.

We began introductions through a novel experience: announcing an interesting or otherwise memorable failure [of our own]. In particular, we were interested in "What have you tried, that just didn't seem to work?" This process set us on even ground: we are equal in our setbacks, and we were there to help each other overcome them.

This seminar followed an *unconference*-style. This means that there was only a very loose structure. There were only three imposed talks, given at the start of the first three days. These talks were given by well-regarded figures (see Section 3, to give broad, deep,

[1] https://nesy-ai.org/
[2] https://neurosymbolic-ai-journal.com/

and historical perspectives of neurosymbolic AI. The remainder of each day was organized into breakouts. These breakout groups were self-selected and even self-generated. After the initial roll-call, we performed a group exercise:

- We wrote down any number of topics that we wanted to tackle this week onto a sticky.
- We placed the topic-sticky onto the chalkboard at the front of the seminar room.
- Small dot stickers were provided for attendees to *upvote* specific topic-stickies.
- We collectively clustered the stickies and identified a theme.

While this reliance on attendee-participation was high-risk, it was also certainly high-reward. However, it was our intuition that the selected participants would be both amenable to this style, but also collegiate in their collective regard for each other, allowing for open, fluid, and vibrant discussion. Indeed, we believe that our risk paid off, and has culminated in this report, as follows.

Overview of the Report

The report is organized into two parts. Section 3 provides the abstracts for the opening talks of the first three days of this seminar. Section 4 contains reports that are jointly written by each breakout group, as described above. Each report provides an overview of the topic, addresses ambitions and challenges, and describes the state of the art.

2 Table of Contents

3 Overview of Opening Talks

For the first three days the seminar, we had an invited talk that would set the stage of Neurosymbolic AI from different perspectives. We provide the abstracts for the latter two. The first talk, given by Frank van Harmelen provided a conversational platform for discussing many of the open problems that our field faces. The second talk, given by Artur d'Avila Gartez, provided both an in-depth discussion on the *technical* aspects of neurosymbolic AI and a historical perspective. Our final invited talk was given by Ute Schmid, who provided context from the cognitive science, as well as a critical look at how knowledge engineering (broadly defined) has underwritten the last four decades of AI research.

3.1 Neurosymbolic AI

Artur d'Avila Garcez (City – University of London, GB)

Artificial Intelligence (AI) has become the focus of large-scale research endeavors in industry and has changed business practice. This led to important debates around the impact of AI on education and society. Concerns around the reliability, fairness, energy efficiency and accountability in AI were raised by influential thinkers [1]. Many identified the need for well-founded knowledge representation and reasoning to be added to the neural-network approach to AI called deep learning. Neurosymbolic AI has been an active area of research for many years seeking to do just that, bringing together robust learning in neural networks with reasoning and symbolic computation.

It has been argued that building AI systems capable of reliable reasoning and safe and trustworthy AI will require neurosymbolic AI systems capable of integrating sound reasoning and deep learning (DL). Parallels have been drawn between Daniel Kahneman's research on human reasoning and decision making [2] and the so-called AI system 1 and system 2. In this keynote, I review the research in neurosymbolic AI and seek to identify promising directions and challenges for the next decade of machine learning (ML) research from the perspective of neurosymbolic computation.

Specifically, I seek to place more than 20 years of research in the area of neurosymbolic AI known as neural-symbolic integration [3] in the context of the recent explosion of interest and excitement around the combination of deep learning and reasoning. I revisit early theoretical results of fundamental relevance to shaping the latest research, such as the proof that recurrently connected, neural networks compute the semantics of various logic formalisms [4]. I also identify bottlenecks and the most promising technical directions in my view towards the sound representation of learning and reasoning in neural networks.

As well as pointing to the various related and promising techniques in neurosymbolic AI, we aim to help organize some of the terminology commonly used around AI, ML and DL. DL is now recognized as being the efficient computational mechanism upon which data-driven AI is to be realized. ML includes DL but also other forms of machine learning such as decision trees, and AI includes machine learning but also reasoning, planning and other abstract cognitive processes. This distinction is important at this exciting time when AI becomes popularized among researchers and practitioners coming from multiple areas of computer science and from other fields altogether, psychology, cognitive science, economics, medicine, engineering and neuroscience.

Recent months have seen a proliferation of releases of Large Language Models by AI companies culminating with the release of GPT5 by ChatGPT's owner OpenAI. Following various announcements of large-scale government and private investments in AI infrastructure, model training and alignment, and heated debates around the safety and risks of AI such as at the World Economic Forum, it has become clear that AI's lack of reliability persists as Large Language Models continue to "hallucinate". The adoption of Agentic AI also failed to solve the problem despite the claims of reduced hallucinations. It turns out that one bad hallucination is sufficient to destroy trust for a long time.

A lot of the evaluation of current AI is based on benchmarking on reasoning tasks with a rather vague definition of reasoning. Neurosymbolic AI has been studying and formalizing reasoning in neural networks for many years [3]. Neurosymbolic AI has as a requirement treating DL as a computational model capable of learning efficiently from data, but also of reasoning logically from what has been learned, incorporating data and Knowledge, formally specified, and satisfying certain verifiable system properties such as correctness.

I survey some of the prominent forms of neural-symbolic integration from the perspective of distributed and localist representations based on the assumption that representation precedes learning. This is possible without having to create a separation between neural and symbolic approaches, as was the motivation of the founding fathers of AI even before the term AI was coined (as in the case of the 1942 paper by McCulloch and Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity, and Von Neumann's 1952 Lectures on Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components, indicating that the gap between distributed vector representations (embeddings) and localist symbolic representations (logic) was not as large as some might imagine. Even Alan Turing's 1948 Intelligent Machinery introduced a type of neural network called a B-type machine). The term Artificial Intelligence was coined ahead of the now famous Dartmouth Workshop, New Hampshire, in 1956. Since then the field has separated into two: symbolic AI and connectionist AI (or neural networks). I argue that this separation into two has delayed progress.

I show how (parts of a) neural network can be coupled with symbolic descriptions with well-defined semantics. I give examples of how propositional, first-order, modal and temporal logic map to and from generative, encoder-decoder and recurrent networks. I illustrate the application of the neurosymbolic cycle (instil and distil knowledge, reason formally, iterate) in a medical diagnosis scenario. The cycle allows domain experts to explain, ask what-if questions and if necessary intervene in the neural network. I argue that the current Chain-of-Thought (CoT) approach to reasoning is misguided in that it tries to improve reasoning by stepping through the input prompts, ignoring the infinite uses of finite means afforded by the combinations of the inputs and its associated accumulation of errors. In neurosymbolic AI, differently from CoT, knowledge is added to neural networks' architectures or loss functions along with a proof of network convergence to stable states. Whether the interpretation is probabilistic or based on many-valued fuzzy logic [5], the neural network learns a probability distribution and its symbolic counterpart provides explainability, knowledge consolidation across multiple tasks, and even extrapolation beyond data distribution. The intended goal is to avoid diverging results and eliminate hallucinations.

References

1 AI Debate 3: The AGI debate, 24 Dec 2022. `https://www.youtube.com/watch?v=JGiLz_Jx9uI`

2 D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

3 A. d'Avila Garcez, K. Broda, D. Gabbay, *Neural-Symbolic Learning Systems: Foundations and Applications*. Springer, New York, 2002.
4 A. d'Avila Garcez, Luis C. Lamb. Neurosymbolic AI: The 3rd Wave. *Artif Intell Rev* **56**, Springer Nature, New York, 2023.
5 S. Odense and A. d'Avila Garcez. A semantic framework for neurosymbolic computation. *Artificial Intelligence* **340**, Elsevier, 2025.

3.2 A Cognitive Perspective on Neurosymbolic AI

Ute Schmid (Universität Bamberg, DE)

Cognitive and AI research mutually inspire each other (very much in early AI, not so much later on, but currently with rising interest again) – giving rise to novel AI methods (e.g., in XAI) as well as influencing cognitive theories. In the talk I first give a reminder on early work on neural learning and symbolic machine learning focusing on relational and rule learning. I will point out crucial differences between neural and symbolic learning with respect to compositionality and productivity, giving illustrations with learning the recursive rule for solving Tower of Hanoi problems and number series induction by pattern abstraction from few examples (rather than generate-and-test). Then I will focus on interleaving implicit and explicit learning with examples from rule-based explanations for blackbox models. I will point out relations to explainable AI for blackbox models. I will discuss abstract visual reasoning as a challenge problem for neurosymbolic approaches and end the talk by pointing out what I think are core challenges for interdisciplinary NeSy research: (1) learning productive rules by abstraction/generalization from few examples, (2) interleaving perception learning and learning complex rules, (3) better understanding of the human inductive bias (generalize the relevant aspects), (4) introducing meta-cognition to control and evaluate generated output (error monitoring, faithfulness), (5) aligning human and machine learning and reasoning for efficient joint problem solving and decision making.

4 Breakout Group Reports

In the following sections, we describe the outcomes of our Breakout Groups. For each outcome report, we specifically ask four questions (in some variation and in some order):

1. What is it?
2. What is the ambition?
3. What are the challenges?
4. Where are we now?

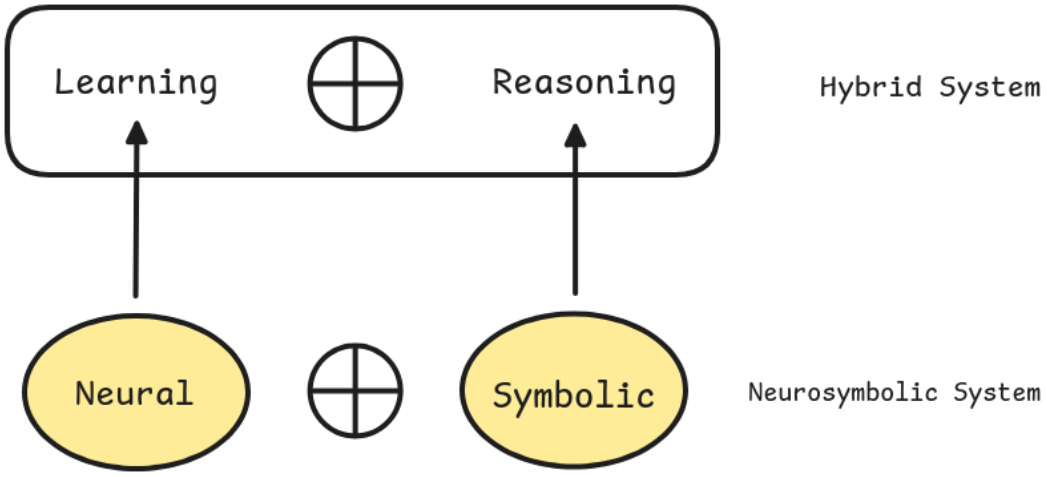


Figure 2 Caption.

4.1 Defining Neurosymbolic Systems

Cogan Shimizu (Wright State University – Dayton, US), Annette ten Teije (VU Amsterdam, NL), Frank van Harmelen (VU Amsterdam, NL)

4.1.1 What are Neurosymbolic Systems?

Neurosymbolic (NeSy) systems aim to unify the strengths of connectionist approaches, such as neural networks, with symbolic reasoning frameworks, such as knowledge graphs and logical inference. This particular variety of hybrid system is motivated by the longstanding observation that neither purely neural nor purely symbolic systems are sufficient to meet the broad demands of general intelligence. While neural networks excel in perception tasks and statistical pattern recognition, they struggle with explainability, data efficiency, and reasoning under constraints. Conversely, symbolic systems offer strong formal guarantees, transparency, and the capacity to integrate explicit knowledge, but are often brittle and lack flexibility in uncertain or noisy environments. This is just one demarcation, which varies in formality, specificity, and perspective.

4.1.1.1 Informal Demarcation

At a high level, neurosymbolic systems operate at the interface of two distinct paradigms, often informally and conveniently characterized along dual axes:

- **System 1 vs. System 2:** fast, unconscious, non-verbal inference (neural) versus slow, deliberate, verbal reasoning (symbolic).
- **Implicit vs. Explicit:** learned representations from data versus human-interpretable rules and knowledge.
- **Black-box vs. Explainable/Interpretable:** opaque computation versus structured reasoning paths.

Yet, for that convenience, we do miss considerable nuance: it is not always the case that either paradigm falls easily or directly onto either axis. Nonetheless, this demarcation

highlights the complementary nature of neural and symbolic components and motivates their integration. A representative description, as articulated by Gartner, defines neurosymbolic AI as "*...a form of composite AI that combines machine learning methods and symbolic systems (for example, knowledge graphs) to create more robust and trustworthy AI models. This combination allows statistical patterns to be combined with explicitly defined rules and knowledge to give AI systems the ability to better represent, reason and generalize concepts. This approach provides a reasoning infrastructure for solving a wider range of business problems more effectively.*"

The key element in this informal demarcation is the combination of a learning component (inferring general patterns from specific instances) and a reasoning component (inferring specific instances from general patterns), as depicted in Figure 2. It does not necessarily hold that the neural (symbolic) component must be the learning (reasoning) component, as there are many ways to compose them or otherwise integrate their properties and functionality.

4.1.1.2 Demarcation & Categorization

While the informal intuition behind neurosymbolic systems is widely shared, precise demarcations remain contested and have changed over time [2, 12]. Key questions include:

- Must the learning component be neural? (e.g., is inductive logic programming [4] included?). This concerns the choice of the learning component, which in the figure has the form of a neural component.
- Must reasoning be symbolic in a strict logical sense? (e.g., is differentiable reasoning [21, 13] in scope? Should LLMs be considered inherently neurosymbolic? Are deep learning systems that solve symbolic tasks neurosymbolic?). This concerns the choice of the reasoning component in the figure.
- How tight should the coupling between the learning and the reasoning component be? This concerns the choice of the $\oplus$ operator in the figure that combines particular learning and reasoning components at the bottom of the figure (e.g., is graph-RAG with an LLM [15] in scope?).
- What constitutes a *symbol*? That is, what is the boundary between symbolic and neural knowledge?

Note that Figure 2 simplifies (for graphical reasons) a neurosymbolic architecture as the combination of a *single* learning component with a *single* reasoning component, while in the literature, more complex combinations have been shown to be useful [10].

Answering these questions is essential for defining the scope of neurosymbolic AI and constructing a taxonomy of system architectures. These include variations in how knowledge is represented in (the architecture [9]), the loss function [5]), the formalism used (fuzzy logic [3]), or probabilistic logic [19]), and the coupling between learning and reasoning components (tight vs. loose integration [14]).

4.1.2 What is the ambition of a theory of NeSy?

The overarching ambition of the neurosymbolic community is to understand and formalize the design space of neurosymbolic systems. While best practices are currently grounded in empirical insights, the field is maturing toward a more principled foundation.

This includes the development of a body of knowledge that:

- Explains why certain architectural choices of components ("neural" n and "symbolic" s in Figure 2) and operators ($\oplus$) work for particular tasks or domains,

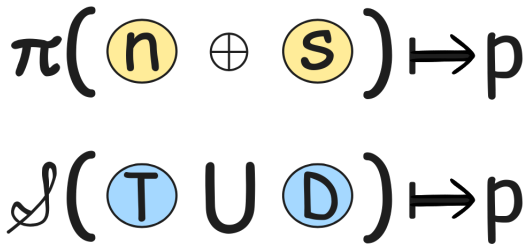


Figure 3 Caption.

- Identifies desirable properties (e.g., safety, robustness, efficiency, scalability, and expressivity) depicted by p in Figure 3.
- Enables formal reasoning about system architecture and behavior, as depicted in Figure 2 as a procedure π that establishes properties p given a specific neurosymbolic architecture.
- Identify tasks and domains (T and D, respectively) for which neurosymbolic approaches would be particularly suitable via $\mathcal{S}$.

In other words, a central goal is to establish a theory of neurosymbolic systems that allows for formal specification, design, and verification. This involves defining input-output behavior, architectural configurations, and desirable properties, and developing proof techniques that ensure certain system properties (safety, efficiency, etc).

4.1.3 What are the challenges for a Theory of NeSy?

The fundamental challenge to a theory of NeSy is the unification of the different views on what NeSy is into a coherent definition, which can be subsequently poured into an appropriate formal form. Such a theory should form the basis for formalisms, methods, and tools for:

- **Specification:** formal descriptions of functional I/O behaviour
- **Design:** formal description of architectures that implement the specification (i.e., the bottom layer in Figure 2).
- **Requirements:** formal definition of desired properties (p in Figure 3).
- **Verification:** proof methods π that derive whether the requirements p hold.

Illustratively, consider a self-driving system deciding whether to stop (specification). A symbolic constraint might enforce that safety conditions must never be violated (requirement). Depending on how these constraints are implemented – embedded in the loss function or enforced as a final reasoning layer (design) – guarantees may differ in strength: a loss function may increase the likelihood of the requirement being met, but will not guarantee it, whereas a final reasoning layer would (verification).

Such a theory of neurosymbolic systems should satisfy the following desiderata, and by their nature, each a challenge unto themselves:

- Descriptive: the theory must describe a broad family of NeSy systems.
- Predictive: the theory must allow us to analytically predict (derive) properties, using operations such as refinement, composition and abstraction, going beyond the purely empirical observations in the current literature.
- Prescriptive: such properties should form the basis for design guidelines that will for the first time advise practitioners which architecture to use for a given set of quality attributes.

- Compositional: the theory would have to be compositional, in the sense that any properties p that the theory derives (π) about a particular neurosymbolic configuration should be a function of the individual components (neural, symbolic) and the way they have been composed ($\oplus$).

4.1.4 Where are we now?

Although individual systems are typically based on solid theoretical grounds (e.g., probabilistic logic for DeepProbLog, Description Logic for Ontolearn, fuzzy logic for LTNs), theories about the entire class of neurosymbolic systems have until now been limited to categorizations of such systems, and have not yet had the shape of a theory as we outlined in two preceding sections. These existing categorizations remain largely informal and only fulfill part of the desired properties from the preceding section:

Table 1 Caption.

Desiderata	**[8]**	**[20]**	**[10]**	**[6]**	**[1]**	**[17, 16]**
Descriptive	+	-	+	-	-	+
Predictive	-	+/-	+/-	+	-	-
Prescriptive	-	+/-	-	+	+	-
Compositional	-	-	-	-	-	+
Formal	-	-	-	+	-	-

Recent developments propose more formal theories of neurosymbolic systems [14, 11], with [14] capturing one of the six patterns proposed in [8]. Although very different in nature, a final noteworthy recent attempt is Uller [18], a Python library for neurosymbolic systems, where (although not intended as such), the abstractions proposed in the library can be seen as "theory" about neurosymbolic architectures.

We see all of these as important precursors for an all-encompassing theory of neurosymbolic systems that is at the same time powerful, has broad coverage, and will be the basis for a well-founded design practice for neurosymbolic systems.

References

1 Elvira Amador-Domínguez, Emilio Serrano, and Daniel Manrique. Neurosymbolic system profiling: A template-based approach. *Knowl. Based Syst.*, 287:111441, 2024.

2 Sebastian Bader and Pascal Hitzler. Dimensions of neural-symbolic integration - A structured survey. In Sergei N. Artëmov, Howard Barringer, Artur S. d'Avila Garcez, Luís C. Lamb, and John Woods, editors, *We Will Show Them! Essays in Honour of Dov Gabbay, Volume One*, pages 167–194. College Publications, 2005.

3 Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

4 Andrew Cropper and Sebastijan Dumancic. Inductive logic programming at 30: a new introduction. *CoRR*, abs/2008.07912, 2020.

5 Claudia d'Amato, Nicola Flavio Quatraro, and Nicola Fanizzi. Injecting background knowledge into embedding models for predictive tasks on knowledge graphs. In Ruben Verborgh, Katja Hose, Heiko Paulheim, Pierre-Antoine Champin, Maria Maleshkova, Óscar Corcho, Petar Ristoski, and Mehwish Alam, editors, *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings*, volume 12731 of *Lecture Notes in Computer Science*, pages 441–457. Springer, 2021.

6 Charles Dickens, Connor Pryor, Changyu Gao, Alon Albalak, Eriq Augustine, William Yang Wang, Stephen J. Wright, and Lise Getoor. A mathematical framework, a taxonomy of

modeling paradigms, and a suite of learning techniques for neural-symbolic systems. *CoRR*, abs/2407.09693, 2024.

7 Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic ai: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, Nov 2023.

8 Henry A. Kautz. The third AI summer: AAAI robert s. engelmore memorial lecture. *AI Mag.*, 43(1):93–104, 2022.

9 Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: neural probabilistic logic programming. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3753–3763, Red Hook, NY, USA, 2018. Curran Associates Inc.

10 Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, 328:104062, 2024.

11 Simon Odense and Artur d'Avila Garcez. A semantic framework for neurosymbolic computation. *Artif. Intell.*, 340:104273, 2025.

12 Md. Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-symbolic artificial intelligence. *AI Commun.*, 34(3):197–209, 2021.

13 Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

14 Lennert De Smet and Luc De Raedt. Defining neurosymbolic AI. *CoRR*, abs/2507.11127, 2025.

15 Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560, 2024.

16 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.

17 Frank van Harmelen and Annette ten Teije. A boxology of design patterns forhybrid learningand reasoning systems. *J. Web Eng.*, 18(1-3):97–124, 2019.

18 Emile van Krieken, Samy Badreddine, Robin Manhaeve, and Eleonora Giunchiglia. ULLER: A unified language for learning and reasoning. In Tarek R. Besold, Artur d'Avila Garcez, Ernesto Jiménez-Ruiz, Roberto Confalonieri, Pranava Madhyastha, and Benedikt Wagner, editors, *Neural-Symbolic Learning and Reasoning - 18th International Conference, NeSy 2024, Barcelona, Spain, September 9-12, 2024, Proceedings, Part I*, volume 14979 of *Lecture Notes in Computer Science*, pages 219–239. Springer, 2024.

19 Emile van Krieken, Thiviyan Thanapalasingam, Jakub M. Tomczak, Frank van Harmelen, and Annette ten Teije. A-nesi: A scalable approximate method for probabilistic neurosymbolic inference. 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.

20 Laura von Rüden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, and Jannis Schuecker. Informed machine learning - A taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.*, 35(1):614–633, 2023.

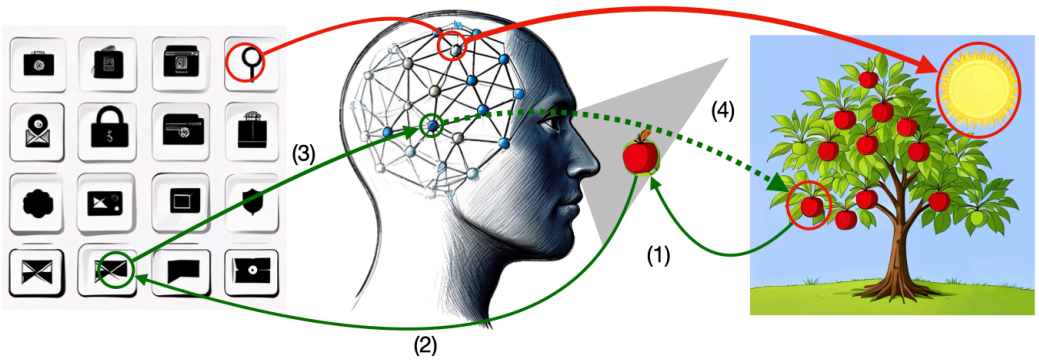


Figure 4 The Symbol Emergence Loop: An agent (1) observes a new concept in the environment, tries to map to an existing symbol, (2) realises that there is a semantic gap with the current symbol systems, and (3) decides to create a new symbol.

21 Po-Wei Wang, Priya L. Donti, Bryan Wilder, and J. Zico Kolter. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. *CoRR*, abs/1905.12149, 2019.

4.2 Symbol Emergence

Riccardo Tommasini (INSA Lyons, FR), Luciano Serafini (Fondazione Bruno Kessler (FBK) – Trento, Italy), Giuseppe Marra (KU Leuven, BE), Jay Pujara (University of Southern California, US), Gustav Šír (Czech Technical University, CZ), Natalia Díaz-Rodríguez (University of Grenada, ES)

4.2.1 What is Symbol Emergence?

This section investigates the conditions under which new symbols should be introduced in the learning process, and how this impacts the semantics and performance of NeSy. Symbol emergence (SE) refers to a symbol being identified from data, gaining recognition as a discrete concept, and becoming useful for reasoning or communicating to one or more agents, according to some pre-defined utility measure. In NeSy, symbols are objects that sit at the boundary between neural and symbolic components, enabling the coupling of these two systems. Symbol emergence provides the seeds for symbol grounding, enabling agents to establish a shared vocabulary for communication.

The symbol emergence problem fills a gap in the "symbol grounding problem" [6], which is defined as endowing agents with the means to autonomously create internal representations that link their manipulated symbols to corresponding referents in the external world. However, the origin of these symbols remains underspecified in the AI community.

4.2.2 What is the NeSy Ambition?

Unlike approaches where symbols are predefined, symbol emergence aims to capture how agents can autonomously form, adapt, and communicate symbolic structures in a manner

that is both grounded in data and useful for reasoning. To sharpen this vision, we suggest some questions:

- What are the desiderata for "good" symbols, and what processes benefit from them?
- How are symbols formed, and what drives their emergence?
- What are the principles according to which symbols emerge?
- What are the processes that underlie symbol creation, adaptation, and forgetting?
- Is a symbol sufficient or necessary for reaching the agent's goal?

Our analytical discussion around these questions led us to formulate the **GRACE^{2}ful** NeSy ambition, a vision for neurosymbolic systems that are generative, robust, adaptable, communicative, efficient, and explainable.

- **Generative (formerly Creative):** Neurosymbolic systems should be capable of inventing new symbols whose semantics are not pre-specified by humans, thereby avoiding human biases. Such creativity goes beyond recombination: it entails analogical reasoning that reuses existing symbolic structures to generate novel conceptualisations. Example: Extending the relation between "apple" and "peel" to reason about "earth" and "crust".
- **Robust:** Emergent symbols must remain stable across variation, noise, and context. A robust NeSys generalises reliably, identifying symbols consistently even under changes in environment or modality. Example: Recognising "apple" whether it appears in sunlight, shadow, or partial occlusion, as well as if it appears in different colours, like in the image above (red and green).
- **Autonomous (formerly Adaptable):** Symbols should not be fixed a priori. Instead, NeSy will continuously refine and expand its symbolic boundaries by autonomously adapting its symbolic repertoire in response to task demands and new input. Example: Introducing "apple" when predicting whether fruits have peels, without prior specification.
- **Communicative:** Symbols must be transferable and composable across agents, humans, or systems, enabling alignment and effective knowledge exchange. Therefore, NeSy will enable knowledge alignment with other systems, agents, or humans by supporting symbols that are appropriately abstracted. For example, a system can provide a description of an apple in either English or French, depending on the user's nationality with whom it is interacting.
- **Efficient:** Emergent symbols should decrease data requirements and computational costs while improving reasoning. For example, a system will be able to use the (emergent) symbol of "apple" to more quickly classify healthy meals. NeSy will benefit from operating on Small Data, reducing reliance on labels, or increasing the efficiency of reasoning.
- **Explainable:** Symbols should make reasoning transparent. Example: Tracing a misclassification of pears as apples back to the boundaries of the symbol "apple". NeSy will be debugged by breaking down processes by their relevant features or sub-processes. For example, mistakes in a prediction can be better understood by inspecting the symbol "apple" and whether it also refers to pears.

4.2.3 What are the challenges, and how to address them?

Realizing the **GRACE^{2}ful** ambition requires overcoming several challenges. Below, we highlight them and relate them to the ambition via Table 2.

- **Criteria for Symbol Emergence.** A central challenge is determining when a new symbol should be introduced. Without clear criteria, systems risk ambiguity, redundancy, or overly narrow symbols. The solution must balance abstraction (generalizing across

Challenge/Property	G	R	A	C	E	E
Criteria for Symbol Emergence	✓		✓			✓
Continuous and Efficient Induction	✓		✓		✓	
Managing Complexity		✓			✓	✓
Semantic Alignment		✓		✓		✓
Shared World Models and ToM		✓	✓	✓		

Table 2 GRACE^2-ful symbols: **G** – generative; **R** – robust; **A** – autonomous; **C** – communicative; **E** – efficient ; **E** – explainable.

trivial variations) with sensitivity (capturing task-relevant distinctions). This challenge is closely tied to the **Generative**, **Autonomous**, and **Explainable** dimensions.
  - *Example:* a fruit recognition system may only need "apple", while a grocery system must distinguish between "Fuji" and "Honeycrisp".
- **Continuous and Efficient Induction.** Beyond criteria, NeSy systems need mechanisms for automatic, continuous, and efficient induction of symbols. Such processes must model symbol semantics (relations between new and existing concepts) and remain scalable. This connects to **Generative**, **Autonomous**, and **Efficient**.
  - *Example:* an agent monitoring sensor streams introduces an "anomaly cluster" when distributions shift.
- **Managing Complexity.** Symbolic reasoning risks combinatorial explosion, where possible symbol combinations grow unmanageably. Systems must introduce symbols that aid efficiency through pruning or prioritisation. This challenge relates to **Robustness**, **Efficiency**, and **Explainability**.
  - *Example:* avoiding fruit–colour–size combinations unless task-relevant.
- **Semantic Alignment.** Emergent symbols must be grounded in perceptual regularities and aligned across two or more agents (who may be humans or machines). Misalignment risks semantic drift. This ties to **Robust**, **Communicative**, and **Explainable**.
  - *Example:* one system uses "apple" only for red apples, another for all apples.
- **Shared World Models and Theory of Mind.** Symbol emergence requires overlapping knowledge or experiences among agents to disambiguate a novel symbol. Indeed, some overlap in experiences is needed to disambiguate potential polysemanticity in symbols. Yet some emergent concepts (textures, scales, properties) resist symbolic capture. This links to **Robust**, **Autonomous**, and **Communicative**.
  - *Example:* two autonomous vehicles must share a compatible notion of 'obstacle'.

4.2.4 Where are we now?

Neurosymbolic systems perform symbol grounding in various ways, including fine-tuning neural networks, clustering, human-guided mapping, or assuming a priori neural-symbolic grounding exists. As NeSy systems incorporate SE, these grounding techniques must be extended to support newly introduced symbols. Several explorations of symbol emergence provide guideposts for designing SE modules in NeSy systems. In this aim, numerous research topics in related fields can benefit NeSy. Taniguchi et al. provide a wide-ranging study of symbol emergence across many fields and consider symbol emergence in robotics [8]. Silver and Mitchell examine the interaction of concepts and symbols in human cognition, utilising evidence from functional magnetic resonance imaging (fMRI) experiments. When examining the literature in Machine Learning, Cognitive Psychology, Social Science, and related fields,

we find a range of techniques closely connected to the problem of symbol emergence (SE). Below, we offer a necessarily incomplete list of approaches from various domains that relate to this issue.

- **Machine learning** has investigated different techniques along which symbolic, sparse or discrete representations arise, independent of possible symbolic processes (reasoning). According to information-theoretic principles, discrete symbols arise from the balance between minimal description length (for compression) and maximal mutual information with the task (for expressiveness).
- When learning **sparse and compressed representations**, guided by principles like information bottleneck and regularisation, naturally promote the emergence of discrete, symbol-like structures by enforcing abstraction and compactness [2, 1].
- **Object-centric learning** aligns with symbol emergence by promoting the disentanglement of scenes into discrete, compositional entities that can be referenced symbolically (e.g., [5]).
- **Causal and disentangled representations** aim to isolate underlying generative factors, which are prime candidates for interpretable and symbolic abstractions [9].
- **Conceptual clustering** partitions continuous data into discrete groups, effectively inducing symbolic categories that can serve as building blocks for higher-level reasoning in neurosymbolic systems. **Incremental clustering** extends clustering methods to cope with data streams that require cluster update (extension/contraction) depending on the data coming in.
- **Structure learning** in the form of program induction or predicate invention (as in Inductive Logic programming) directly engages with symbolic representations, formalizing the emergence of new symbols and relations from data. Symbols can emerge through interaction with humans or other agents, as shared meanings are gradually negotiated and grounded in communicative behaviour and mutual experience [4, 7].
- **Architectural biases** in machine learning models, such as discrete relaxations (Gumbel-softmax [3]) or sparsity constraints, encourage neural models to adopt symbolic-like behavior, facilitating differentiable approximations of symbolic reasoning.

References

1 Javier Blanco-Romero, Vicente Lorenzo, Florina Almenares Mendoza, and Daniel Díaz-Sánchez. Machine learning predictors for min-entropy estimation, 2024.

2 Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

3 Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.

4 Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*, 2017.

5 Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

6 Dairon Rodríguez, Jorge Hermosillo, and Bruno Lara. Meaning in artificial agents: The symbol grounding problem revisited. *Minds Mach.*, 22(1):25–34, 2012.
7 Luc Steels. Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312, 2003.
8 Tadahiro Taniguchi, Justus H. Piater, Florentin Wörgötter, Emre Ugur, Matej Hoffmann, Lorenzo Jamone, Takayuki Nagai, Benjamin Rosman, Toshihiko Matsuka, Naoto Iwahashi, and Erhan Öztop. Symbol emergence in cognitive developmental systems: A survey. *IEEE Trans. Cogn. Dev. Syst.*, 11(4):494–516, 2019.
9 Kevin Xia and Elias Bareinboim. Neural causal abstractions. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 20585–20595. AAAI Press, 2024.

4.3 Small Data and Neurosymbolic AI

Filip Ilievski (VU Amsterdam, NL), Axel-Cyrille Ngonga Ngomo (University of Paderborn, DE), Hande McGinty (Kansas State University – Manhattan, US), Valentina Tamma (University of Liverpool, UK)

4.3.1 What is it?

This section assumes NeSy systems applied to supervised machine learning. In this setting, the systems are provided with a set of training data items $D \subseteq X \times Y$, where X is drawn from the set of all possible inputs U, and Y is the corresponding output. The systems then aim to approximate an optimal target function $f : U \to Y$ by learning $\hat{f} : U \to Y$. The performance of $\hat{f}$ is finally measured in terms of bounded performance measures such as accuracy, F-measure, mean reciprocal rank, and hits@n. Within supervised machine learning, small data is commonly understood in terms of limited training data in terms of **size** [7, 26, 60] or **statistical moments** prior to training [57]. While other definitions related to data models are also found in the literature [58, 48], these go beyond the scope of this section.

The first interpretation of small data is often seen as a counterpart to the large amount of data needed to train deep learning systems. For example, transformer-based large language models are routinely trained with $> 10^{12}$ tokens before being aligned using thousands of annotated data samples [8]. Given that the provision of such samples is often costly, the ability of systems to achieve high performance after having been trained on limited magnitude data sets D is aimed at reducing annotation time and costs, training time, and even improving energy efficiency. The second definition of small data is correlated with – but not equivalent to – the first one and refers to data with a small data statistic T (e.g., entropy) in comparison to the dimensionality of the problem [57]. This definition equates to regarding data as small if the distribution of samples within or across classes is skewed with respect to some statistical moment (e.g., frequency, standard deviation). Such limitations are predominant when data annotation costs are high (for example, in medicine, where image annotation costs can reach $50 USD/image) or when data availability is limited before training (e.g., when building industrial plants).

4.3.2 What is the NeSy Ambition?

NeSy systems seem particularly well suited for being trained with small data (according to both meanings of the term) as they promise to address a key limitation of purely data-driven systems: to limit their dependence on training data to discriminate across instance-class pairs by an explicit use of symbolic (and subsymbolic) background knowledge. Specifically, NeSy AI aims to support small-data learning, reasoning, and evaluation. First, NeSy AI aims to **learn generalizable features** from the data, including implicit features captured in the subsymbolic space (e.g., prototype neurons [22]), concepts (e.g., learning with less than 10 samples for classification tasks [10, 46, 37, 13]), and explicit symbolic features that are meaningful to humans (e.g., using object parts to recognize sketches in computer vision [9]). Meaningful decision-making requires that the features in the symbolic and subsymbolic space (i.e., neurons and symbols) should be at least alignable [50, 43]. Second, NeSy AI promises to enable **knowledge-enriched explicit symbolic reasoning** over the derived symbolic features. Since small data is, in general, unlikely to be sufficient for models to learn how to solve the task at hand, abstracting to symbols provides an opportunity for symbolic reasoning and the incorporation of contextual commonsense and causal knowledge [28]. Third, NeSy AI aims to **develop principled procedures for evaluating small data** use cases. NeSy frameworks can provide meaningful auditing of systems in line with quality attributes [56]. The focus on symbolic representations enables the adoption and adaptation of cognitive mechanisms, such as prototypes, analogy, and rule learning, for benchmarking purposes [29]. Carefully designed human intelligence tests have already inspired a long list of AI benchmarks, such as those in abstract visual reasoning [24, 40]. Conversely, the strong link of NeSy to machine learning enables the introduction of statistical methods that can support meaningful evaluation, such as sampling methods that derive subsets from existing datasets that fit specific statistical properties and possibly simulate long-tail phenomena [47].

Drawing on SotA methods for learning, reasoning, and evaluation makes the NeSy methods a promising fit for a variety of **applications where data collection may be laborious, expensive, or dangerous**. This is apparent for sensitive applications where data collection can be challenging and potentially dangerous to one's health, like medical trials [1, 12]; applications with privacy, security, and ethics concerns like molecular science [15]; and applications where the high cost of error hinders large-scale data acquisition, such as autonomous driving [21, 27]. More broadly, NeSy methods aim to support any task in domains with high irregularities that heavily rely on causal and what-if reasoning, such as traffic [59] and ecology [54] domains. Advancements in learning, reasoning, and evaluation in NeSy methods can improve data efficiency, interpretability, and safety in robotic domains, especially for tasks that require embodied decision-making and structured planning from limited demonstrations [55].

4.3.3 What are the challenges and their mitigations?

When working with small datasets, a fundamental challenge arises from their **limited depth and breadth**. Small data often fails to capture the full range of possible variations and alternative perspectives in a domain. For example, clinical trial data may have some variability within a cohort, it is unlikely that the data set is representative of the demographic diversity, potential side effects, or rare complications. As these datasets cannot be easily extended to include what is missing, injecting background knowledge can help bridge gaps and enrich the dataset with relevant contextual information that is otherwise unavailable. However, this process underscores the necessity of adequate abstractions that preserve essential patterns

while generalizing beyond the sparse data available. Without careful abstraction, the utility of both the limited dataset and the supplementary knowledge may be diminished. Key challenges are applying the most **appropriate abstraction**, the need to **incorporate discrete and continuous representations** (e.g., knowledge graphs into neural networks), and ensuring **trustworthy background knowledge** (e.g., using provenance to compensate for absent or underrepresented aspects of the data).

Various NeSy methods offer a trade-off between the richness of prior knowledge and the burden of integrating it (§4.3.4). *Prototypes* are an approach to abstraction, with a question of where they come from: they can be specified by experts, which is relatively cheap, or learned from data, which requires sufficient coverage and well-structured input. Class *hierarchies* can guide the learning of prototypes [22], but it remains uncertain how well these approaches perform under truly small data constraints. *Ontology building* through middle-out approaches [44, 53, 5, 42, 49, 25], where the class definition is refined to cover some sample data that domain experts provide, may enable prototype learning and support flexible abstraction. While *case-based reasoning* can generalize to training data samples, a key challenge is representing the cases at the right level of granularity/abstraction and devising a corresponding similarity metric [23]. *Causal relationships* can inform generalization in new situations, such as using drug side effects data to infer multiple drug-drug interactions. While causal knowledge injection has strong generalization power, especially for counterfactual or rare-event reasoning, causal knowledge is challenging to obtain and validate, requiring strong assumptions. Across the approaches, a clear regularity emerges: to compensate for limited labeled data, we inject additional knowledge into the learning process, whether in the form of abstractions, background knowledge, prototypes, cases, or constraints, which invariably carry costs: acquiring knowledge, representing it effectively, and ensuring its relevance are non-trivial challenges.

Another increasingly recognized challenge concerns evaluation. Conventional performance metrics, such as accuracy, precision, recall, and F_1 score, enable well-understood ground for comparison, but do not account for data size and can lead to misleading assessments of model reliability in low-resource settings. In the context of NeSy systems, this issue is particularly significant given the interplay between symbolic reasoning and statistical learning. To address this, there is a growing need for evaluation **metrics that explicitly consider the size and distributional characteristics of the training data**. These might include data-aware performance bounds, sample-efficiency metrics, or metrics that incorporate statistical moments. However, while performance-related metrics are relatively well-established, other vital **properties of NeSy systems, such as safety, robustness, and explainability, remain challenging to quantify**. Existing proxy measures for explainability (e.g., fidelity, sparsity, or concept alignment) are still far from standardized and often fail to capture user-centered or domain-specific needs [2]. Similarly, there is **a lack of principled and widely adopted metrics** for evaluating the safety of model outputs, particularly in distribution shift or in high-stakes decision-making contexts. Developing rigorous, interpretable, and context-sensitive metrics for these properties remains an open and pressing challenge in the field.

4.3.4 Where are we now?

NeSy AI approaches have shown promise in learning generalizable features, reasoning over them using knowledge-enriched symbolic methods, and enabling principled evaluation tailored to these capabilities. The integration of symbolic structures (e.g., logic rules, background knowledge, compositional constraints) into neural learning processes can enhance the inherent

semantics of limited or redundant data. Structured priors help constrain the hypothesis space, guide generalization, and support abstraction, all of which are critical in data-scarce and low-diversity settings. We review SotA benchmarks and methods next.

4.3.4.1 Benchmark Datasets and Tasks for Small Data

Benchmarks are essential to evaluate and drive progress in NeSy methods, especially under small-data or low-diversity conditions. Notable examples include:

1. **Abstract visual reasoning benchmarks** [24, 40]
   a. *Abstraction and Reasoning Corpus (ARC)* [11]: ARC focuses on abstract patterns and limited training sets. As such, ARC poses challenges that naturally align with NeSy approaches due to its emphasis on compositionality and generalization from minimal data.
   b. *Raven's Progressive Matrices (RPMs)* [45, 3] test abstract visual reasoning through matrix pattern completion and has been widely adopted to study systematic generalization. NeSy methods excel here by incorporating explicit relational and logical reasoning.
   c. *Bongard Problems* [6]: Classic tests of concept learning and analogical reasoning, Bongard problems require understanding subtle, often symbolic, distinctions from very few examples, making them a canonical challenge for neurosymbolic AI.
   d. *MARVEL* [31]: designed to generalize abstract visual reasoning tasks and to separate perception and inference explicitly. MARVEL tests the ability to learn from limited visual data while leveraging symbolic reasoning modules.
2. **Sketch recognition tasks** [39, 18]: While sometimes debated as a small-data problem, sketch recognition benchmarks involving human-like sparse sketches can require efficient abstraction and symbolic interpretation.
3. **Concept learning benchmarks** [35]: Tasks that test the ability to learn symbolic concepts from few examples, often requiring compositional generalization and abstraction capabilities.
4. **Lateral thinking puzzles** [30, 34]: Tasks that test the ability of models to overwrite commonsense associations, by the virtue of adequate abstractions in textual brain teasers and multimodal rebus puzzles.
5. **Structural analogies** [52, 4] are important drivers of knowledge transfer in humans. Recent work has introduced benchmarks for analogies in structures, specifically in images and stories, inspired by cognitive science theories of analogy like structural mapping [20].

4.3.4.2 Methods for Small Data Scenarios

In both small and low-diversity data regimes, symbolic structures serve as a form of *semantic injection*: providing constraints, abstractions, or prior knowledge that compensates for the limitations of the raw data. Several promising categories of NeSy methods that inject symbolic structures have emerged (summarized in Table 3):

1. **Inductive logic programming (ILP) and differentiable reasoning**: Differentiable ILP frameworks and neural theorem-provers can learn logical rules from limited datasets, particularly when symbolic background knowledge is encoded [17]. These systems benefit from structured search spaces and gradient-based learning to optimize over logical rules.
2. **Transfer learning in symbolic contexts**: In transfer learning, neural networks pre-trained on large corpora are fine-tuned with limited task-specific data. Systems such as DeepProbLog [41] integrate neural predicates into logic programs, allowing gradients to

flow through symbolic reasoning pipelines. Domain-adversarial training [19] can further enhance generalization when no labeled data are available in the target domain.

3. **Prototype-based learning and concept bottlenecks**: Few-shot models such as Prototypical Networks [51] summarize each class with a prototype in latent space. When guided by symbolic descriptors (e.g., logic-derived attributes), they enforce structure and boost generalization. Similarly, Concept Bottleneck Models [33] use human-defined concepts to align internal representations, reducing sample complexity and improving interpretability, especially with low-diversity data.
4. **Neurosymbolic case-based reasoning (CBR)** systems combine neural encoders with symbolic memory or prototype mechanisms [38]. By comparing new inputs with stored prototypes or cases, they improve generalization from few or homogeneous examples, often using auto-encoders or prototype networks to learn structured latent representations.
5. **Knowledge injection and logic-based regularization**: Models like Logic Tensor Networks (LTNs) [14] and DeepProbLog [41] integrate logical constraints and knowledge bases into learning objectives, while others inject background knowledge to guide embedding representations [16] or adapt LLMs [28]. These priors regularize training and improve robustness under data sparsity or low variability.
6. **Causal and compositional generalization**: NeSy models increasingly target abstract reasoning, compositionality, and causality, which are key capacities for robust generalization from limited or redundant data [36]. Structural inductive biases (e.g., causal graphs, symbolic decompositions) enhance learning efficiency in such regimes. Entailment trees, that may be modeled as belief graphs [32], further support learning under small-data conditions by enforcing consistency and structure in reasoning. By explicitly modeling inferential dependencies between symbolic statements, they enable the system to generalize from sparse evidence through accurate, compositional inference.

References

1 Scott Askin, Denis Burkhalter, Gilda Calado, and Samar El Dakrouni. Artificial intelligence applied to clinical trials: opportunities and challenges. *Health and technology*, 13(2):203–213, 2023.

2 Roberto Barile, Claudia d'Amato, and Nicola Fanizzi. Lp-dixit: Evaluating explanations for link predictions on knowledge graphs using large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 4034–4042, 2025.

3 David G T Barrett, Felix Hill, Adam Santoro, Ari S Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. *International Conference on Machine Learning (ICML)*, 2018.

4 Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. VASR: visual analogies of situation recognition. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

5 Eva Blomqvist, Karl Hammar, and Valentina Presutti. Engineering ontologies with patterns–the extreme design methodology. In *Ontology Engineering with Ontology Design Patterns*, pages 23–50. IOS Press, 2016.

6 M. M. Bongard. *Pattern Recognition*. Herzen State Pedagogical Institute, 1970.

7 Lorenzo Brigato and Luca Iocchi. A close look at deep learning with small data. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2490–2497. IEEE, 2021.

8 Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered:

Table 3 Expanded summary of NeSy approaches and their mechanisms for learning from small data.

Approach	Symbolic Component	Neural Component	Small Data Advantage
1. Inductive Logic Programming (ILP) and Differentiable Reasoning			
Inductive Logic Programming (ILP)	Logical rules, background knowledge, Prolog-style programs	Differentiable rule scoring, logic unfolding via neural models	Reduces hypothesis space and enables rule induction from few examples
Differentiable ILP / Neural Theorem Proving	Soft logic operators, logic programs	End-to-end differentiable optimization over rule space	Enables gradient-based learning of logical structures under weak supervision
Logic Tensor Networks (LTN) / Logic-Based Regularization	First-order logic rules translated into differentiable constraints	Neural networks trained under logic-derived loss terms	Encodes symbolic priors as soft constraints, improving sample efficiency
2. Transfer and Probabilistic Logic Models			
DeepProbLog / NeSy Probabilistic Logic	Probabilistic logic programs with symbolic predicates and rules	Neural modules embedded as learnable predicates	Allows few-shot or zero-shot learning by leveraging symbolic logic during training and inference
Domain Adaptation / Transfer Learning with Symbolic Regularization	Source-target mappings, task-level symbolic structure	Domain-invariant neural representations, often adversarially trained	Transfers pretrained knowledge with minimal labeled target data by aligning feature spaces
3. Prototype and Concept-Based Models			
Prototype Networks	Optionally logic-guided feature selection, prototype semantics	Latent space embedding and prototype-based metric learning	Learns class-specific prototypes that support generalization from few labeled examples
Concept Bottleneck Models (CBM)	Concept taxonomy or human-annotated concept space	Neural concept predictors with interpretable bottlenecks	Supervizes intermediate representations, improving generalization and interpretability
Neurosymbolic Case-Based Reasoning (CBR)	Symbolic case library or prototype memory, similarity metrics	Neural encoder-decoder or autoencoder with prototype layer	Enables retrieval-based generalization and reconstruction from small training sets
4. Program Synthesis and Neuro-Guided Search			
Program Synthesis (e.g., DeepCoder, DreamCoder)	Domain-specific language (DSL), symbolic grammars, type constraints	Neural models that guide search or program induction	Solves tasks with few I/O examples by learning efficient symbolic programs
5. Causal and Compositional Generalization			
Causal and Compositional Reasoning Models	Causal graphs, compositional rule structures, abstract hierarchies	Modular neural architectures or neural-symbolic hybrids	Supports abstraction and robust generalization under distributional shifts or novel combinations

A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 2024.

9 Kezhen Chen, Kenneth D Forbus, Balaji Vasan Srinivasan, Niyati Chhaya, and Madeline Usher. Sketch recognition via part-based hierarchical analogical learning. In *IJCAI*, pages 2967–2974, 2023.

10 Kezhen Chen, Irina Rabkina, Matthew D McLure, and Kenneth D Forbus. Human-like sketch object recognition via analogical learning. In *Proceedings of the AAAI Conference*

on Artificial Intelligence, volume 33, pages 1336–1343, 2019.

11 François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

12 Hitesh Chopra, Dong K Shin, Kavita Munjal, Kuldeep Dhama, Talha B Emran, et al. Revolutionizing clinical trials: the role of ai in accelerating medical breakthroughs. *International Journal of Surgery*, 109(12):4211–4220, 2023.

13 Caglar Demir and Axel-Cyrille Ngonga Ngomo. Neuro-symbolic class expression learning. In *IJCAI*, pages 3624–3632, 2023.

14 Ivan Donadello, Luciano Serafini, and Artur d'Avila Garcez. Logic tensor networks for semantic image interpretation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1596–1602, 2017.

15 Bozheng Dou, Zailiang Zhu, Ekaterina Merkurjev, Lu Ke, Long Chen, Jian Jiang, Yueying Zhu, Jie Liu, Bengong Zhang, and Guo-Wei Wei. Machine learning methods for small data challenges in molecular science. *Chemical Reviews*, 123(13):8736–8780, 2023.

16 Claudia d'Amato, Nicola Flavio Quatraro, and Nicola Fanizzi. Injecting background knowledge into embedding models for predictive tasks on knowledge graphs. In *European Semantic Web Conference*, pages 441–457. Springer, 2021.

17 Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

18 Yan Fu, Yibing Xu, Yanning Chen, and Alan L. Yuille. Learning to recognize sketches by spatio-temporal graph convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):3007–3019, 2018.

19 Yaroslav Ganin and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

20 Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

21 Leilani H Gilpin and Filip Ilievski. Neuro-symbolic reasoning in the traffic domain. *J AI Res*, 15(3):123–145, 2021.

22 Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 32–40, 2019.

23 Jerry Zhi-Yang He, Zackory Erickson, Daniel S Brown, Aditi Raghunathan, and Anca Dragan. Learning representations that enable generalization in assistive tasks. In *Conference on Robot Learning*, pages 2105–2114. PMLR, 2023.

24 José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L. Dowe. Computer models solving intelligence test problems: Progress and implications (extended abstract). In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 5005–5009. ijcai.org, 2017.

25 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

26 Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

27 Atalay Mert Ileri, Nalen Rangarajan, Jack Cannell, and Hande McGinty. Vel: A formally verified reasoner for owl2 el profile, 2024.

28 Filip Ilievski. Human-centric ai with common sense, 2024.

29 Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, et al.

Aligning generalisation between humans and machines. *arXiv preprint arXiv:2411.15626*, 2024.

30 Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. Brainteaser: Lateral thinking puzzles for large language models. *EMNLP*, 2023.

31 Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 46567–46592. Curran Associates, Inc., 2024.

32 Nora Kassner, Øyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schütze, and Peter Clark. Language models with rationality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14190–14201. Association for Computational Linguistics, 2023.

33 Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

34 Koen Kraaijveld, Yifan Jiang, Kaixin Ma, and Filip Ilievski. COLUMBUS: Evaluating cognitive lateral understanding through multiple-choice rebuses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4410–4418, 2025.

35 Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

36 Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

37 Jens Lehmann, Sören Auer, Lorenz Bühmann, and Sebastian Tramp. Class expression learning for ontology engineering. *Journal of Web Semantics*, 9(1):71–81, 2011.

38 Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

39 Marcus Liwicki and Horst Bunke. A novel approach to on-line handwriting recognition based on combined symbol and structural information. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 211–216. IEEE, 2005.

40 Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736, 2023.

41 Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31, pages 3749–3759, 2018.

42 Hande Küçük McGinty. *KNowledge Acquisition and Representation Methodology (KNARM) and its applications*. PhD thesis, University of Miami, 2018.

43 Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*, 2024.

44 Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

45 John C. Raven. *Raven manual: Section 3: Standard progressive matrices*. Pearson, 1998.

46 Giuseppe Rizzo, Nicola Fanizzi, and Claudia d'Amato. Class expression induction as concept space exploration: From dl-foil to dl-focl. *Future Generation Computer Systems*, 108:256–272, 2020.

47 Michael Röder, Pham Thuy Sy Nguyen, Felix Conrads, Ana Alexandra Morim da Silva, and Axel-Cyrille Ngonga Ngomo. Lemming - example-based mimicking of knowledge graphs. In *15th IEEE International Conference on Semantic Computing, ICSC 2021, Laguna Hills, CA, USA, January 27-29, 2021*, pages 62–69. IEEE, 2021.

48 Hadas Shalit Peleg and Anat Milo. Small data can play a big role in chemical discovery. *Angewandte Chemie*, 135(26):e202219070, 2023.

49 Cogan Shimizu, Karl Hammar, and Pascal Hitzler. Modular ontology modeling. *Semantic Web*, 14(3):459–489, 2023.

50 Daniel L. Silver and Tom M. Mitchell. The roles of symbols in neural-based ai: They are not what you think!, 2023.

51 Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4077–4087, 2017.

52 Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. Arn: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086, 2024.

53 Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The neon methodology for ontology engineering. In *Ontology engineering in a networked world*, pages 9–34. Springer, 2011.

54 William J. Sutherland, Jake M. Robinson, David C. Aldridge, tim Alamenciak, Matthew Armes, Nina Baranduin, Andrew J. Bladon, Martin F. Breed, Nicki Dyas, Chris S. Elphick, Richard A. Griffiths, Jonny Hughes, Beccy Middleton, Nick A. Littlewood, Roger Mitchell, William H. Morgan, Roy Mosley, Silviu O. Petrovan, Kit Prendergast, Euan G. Ritchie, Hugh Raven, Rebecca K. Smith, Sarah H. Watts, and Ann Thornton. Editorial: Creating testable questions in practical conservation: a process and 100 questions. *Conservation Evidence Journal*, 19, Jan 2022.

55 Emre Ugur, Alper Ahmetoglu, Yukie Nagai, Tadahiro Taniguchi, Matteo Saveriano, and Erhan Oztop. Neuro-symbolic robotics. *IEEE Transactions on Robotics (review article)*, 2025. Comprehensive review of neural–symbolic integration in robotic architectures.

56 Frank Van Harmelen and Annette Ten Teije. A boxology of design patterns for hybrid learning and reasoning systems. *Journal of Web Engineering*, 18(1-3):97–123, 2019.

57 Edoardo Vecchi, Lukás Pospísil, Steffen Albrecht, Terence J. O'Kane, and Illia Horenko. espa+: Scalable entropy-optimal machine learning classification for small data problems. *Neural Comput.*, 34(5):1220–1255, 2022.

58 Alvaro Velasquez, Neel Bhatt, Ufuk Topcu, Zhangyang Wang, Katia Sycara, Simon Stepputtis, Sandeep Neema, and Gautam Vallabha. Neurosymbolic ai as an antithesis to scaling laws. *PNAS Nexus*, 4(5):pgaf117, 05 2025.

59 Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9878–9888, 2021.

60 Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9(1):42, 2023.

4.4 Explainable AI

Pascal Hitzler (Kansas State University – Manhattan, US), Catia Pesquita (Universidade de Lisboa, PT), Michael L. Raymer (Wright State University – Dayton, US), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Daria Stepanova (Bosch Center for AI, DE)

4.4.1 What is Explainable AI?

Explainable AI (XAI) [9, 27, 1, 5, 20, 25, 26, 2] aims to communicate information about an AI system that can help to understand or assess validity of system output. The need for XAI arrives out of the fact that AI systems such as those based purely on deep learning are often "black boxes" with obscure internal mechanisms that do not readily allow for an understanding how certain inputs lead to certain outputs.

Explanations in XAI can be characterized along several interrelated dimensions, each shaping the nature of the explanation, its integration with the underlying model, and its utility for various stakeholders. While these dimensions are not entirely independent, they offer a valuable conceptual framework for the design of explanation techniques. One critical consideration is the purpose of the explanation (see e.g., [6]) – whether it is intended to expose the internal structure and operation of the model, to support debugging and refinement, to justify or validate model outputs in a way that fosters trust and confidence, or to offer insights into the underlying data-generating process or real-world phenomena.

The form that an explanation takes may also vary, ranging from machine-readable outputs that enable model introspection, comparison, or reasoning, to human-interpretable formats such as visualizations (e.g., heatmaps or causal graphs) or natural language descriptions. The intended audience further shapes the design of an explanation, which might be tailored for domain experts, lay users, model developers, automated agents, or regulatory bodies. Additionally, explanations may differ in their relationship to the model itself: some systems are explicitly designed to produce inherently interpretable outputs [11, 22], while others rely on post-hoc techniques to extract explanations from otherwise opaque reasoning processes [8, 28, 7, 12, 18, 17, 16]. Finally, explanations may convey meaning at varying levels of abstraction. While low-level, granular accounts (e.g., visualizations of activation patterns or weight dynamics) can offer insight into the internal mechanics of a model, more meaningful and actionable explanations often emerge at higher levels of abstraction (see also Section 4.2). This is analogous to thermodynamics, where macroscopic properties like temperature and pressure provide more interpretable and practically relevant information than the exhaustive specification of each particle's motion and energy.

4.4.2 What is the NeSy Ambition?

Logic-based AI systems are inherently explainable to some extent, e.g., they allow for the examination of reasoning chains (like proof trees), identification of key facts influencing outputs (like abductive reasoning), and the tracing of conclusions back to foundational assumptions (such as whether the system operates under an open or closed world view). Neurosymbolic AI systems that are hybrid (i.e., consist of coupled neural and symbolic systems) are therefore inherently partially explainable because their symbolic components are.

In contrast, post-hoc explanation methods are used after an AI system (say, one based on deep learning) has been trained. Neurosymbolic post-hoc methods typically produce a set of logical axioms that may (partially) capture the network's input-output behavior, and/or internal activation propagation (say, in the form of logical rules [28, 7], and/or labels for hidden node activation patterns (say, as description logic concepts [8]). The resulting logical axioms can then be used for additional reasoning or analysis, such as detecting contradictions, exploring inference paths, model improvements, bias detection, or verifying assumptions, ultimately improving quality of output and trust in the system, for the end user or within collaborative human-AI teaming efforts [14]. Explanations that take hidden node activations and their propagation into account can also help ground the explanation in the actual run-time mechanics of the neural model, increasing trust and interpretability. Post-hoc explanations can also bring in implicit knowledge that is not part of the task input [13].

Neurosymbolic explainability can enable domain experts – like a biochemist – to understand underlying mechanisms, such as identifying biomarkers for a disease. Through its use of logic-based knowledge representation, it also opens up XAI to to full power and versatility of formal semantics. As such, neurosymbolic approaches to XAI can also draw from external knowledge in order to provide independently verifiable evidence for AI claims in responses; a possible pathway to establish external validity of post-hoc rationalizations. Ultimately, hybrid NeSy supports not just explainable models, but systems that invite deeper human understanding and interaction.

4.4.3 What are the challenges?

There are several challenges to fulfilling the NeSy ambition that are rooted in the overarching challenges of explainability [23] but are tied to the particulars of NeSy.

One major issue lies in ensuring the actual usefulness of explanations [4]: while NeSy methods are able to generate explanations that are both faithful to the model's output and coherent with the knowledge model, they may not serve the user's needs or context. In fact, current research in explainable AI has faced criticism for being driven by researchers' assumptions rather than the actual needs of end users, often overlooking who the explanations are truly for [23]. As a result, there is growing momentum toward adopting a human-centered approach in XAI to ensure explanations are meaningful and tailored to specific stakeholders [20, 21].

This misalignment often stems from mismatched expectations around explanatory depth, abstraction level, and complexity – users may receive explanations that are either too general to be informative, too detailed, or too large to be understandable. Balancing depth and simplicity becomes crucial, particularly as different users (e.g., experts vs. laypersons) require different levels of abstraction. A clear example would be an explanation that lacks relevance, e.g., when explaining a specific drug recommendation (e.g., sunitinib[3]) for a cancer patient, a faithful, coherent, and useless explanation would be (patient -has→ cancer -is treated by→ anti-cancer drugs ← is a - sunitinib). This explanation does not afford sufficient explanatory depth to fulfill the purpose of validating the recommendation and supporting a medical decision. On the other hand, explanations at a similar abstraction level may not serve the specific user's purpose. For example, while the following could be an appropriate explanation targeting a medical doctor: patient -has-diagnosis→ clear cell renal carcinoma -has-guideline-treatment→ sunitinib; it would not suit the purposes of a drug development researcher, who likely requires an explanation focusing on the molecular mechanisms.

[3] https://en.wikipedia.org/wiki/Sunitinib

In turn, this also introduces the challenge of the intelligibility of an explanation, i.e., a symbolic explanation may be faithful, coherent, and relevant, but not be comprehensible to the user. Symbolic reasoning chains are indeed not automatically intelligible; translating them into a communicable format (be it natural language or a diagram, etc) that reflects human explanatory norms is challenging, particularly if the symbols do not map intuitively to a user's understanding (re. symbol grounding problem).

Additional challenges are introduced by the incompleteness or outdatedness of the knowledge model [4]. When domain knowledge evolves quickly, explanations might lag behind, giving rise to temporal drift/timeliness by relying on obsolete or potentially incorrect information. Moreover, NeSy explanations should be able to gracefully manage the tension between adhering to established knowledge and incorporating novel patterns that contradict it, to reveal new insights, supporting an unconfined exploration/dynamic-scoped exploration.

NeSy explainability also faces the transversal issue of scalability. As the knowledge base grows or the domain evolves rapidly – as in healthcare or finance – maintaining timely, relevant, and accurate explanations becomes increasingly difficult. Real-time symbolic reasoning is computationally intensive, especially when personalized or context-sensitive explanations are required. In dynamic domains, outdated knowledge can lead to stale or misleading explanations unless the system can adapt to temporal changes. Furthermore, as NeSy systems are deployed more broadly, they must support a diverse range of users with different informational needs and levels of expertise. This necessitates scalable user modeling and context-aware explanation strategies. Ultimately, the promise of NeSy explainability hinges not just on fidelity to reasoning processes, but on the system's ability to communicate those processes effectively and adaptively to humans.

In order to make progress on XAI, concerted benchmarking and evaluation efforts are also required, that should take the above mentioned aspects into account. At this point in time high-quality benchmarks remain to be established for this purpose (see also Section 4.8).

4.4.4 Where are we now?

Research into neurosymbolic explainable AI within hybrid neurosymbolic systems is underway, with promising early work across several key areas [11, 22, 8, 28, 7, 12, 14, 4, 15, 10, 3, 24, 5, 19]), some of which we already touched upon above. However the state of the field is clearly still exploratory, in that a multitude of methods is currently being proposed and tried out. And while individual approaches show promise, they need to be rigorously validated, better integrated into coherent systems, and evaluated through strong proof-of-concept implementations. Crucially, future work must demonstrate how these approaches enhance decision quality and foster user trust.

References

1 Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

2 Sajid Ali, Tamer Abuhmed, Shaker H. Ali El-Sappagh, Khan Muhammad, Jose Maria Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz Rodríguez, and Francisco Herrera. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion*, 99:101805, 2023.

3 Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frédéric Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In Andreas Krause,

Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 1801–1825. PMLR, 2023.

4 Roberto Barile, Claudia d'Amato, and Nicola Fanizzi. LP-DIXIT: evaluating explanations for link predictions on knowledge graphs using large language models. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov, editors, *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 4034–4042. ACM, 2025.

5 Adrien Bennetot, Gianni Franchi, Javier Del Ser, Raja Chatila, and Natalia Díaz Rodríguez. Greybox XAI: A neural-symbolic learning framework to produce interpretable predictions for image classification. *Knowl. Based Syst.*, 258:109947, 2022.

6 David J. Chalmers. Propositional interpretability in artificial intelligence. *CoRR*, abs/2501.15740, 2025.

7 Gabriele Ciravegna, Francesco Giannini, Pietro Barbiero, Marco Gori, Pietro Lio, Marco Maggini, and Stefano Melacci. Learning logic explanations by neural networks. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 547–558. IOS Press, 2023.

8 Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Samatha Ereshi Akkamahadevi, Avishek Das, Cara Widmer, Eugene Y. Vasserman, Md Kamruzzaman Sarker, and Pascal Hitzler. Towards a neurosymbolic understanding of hidden neuron activations. *Neurosymbolic Artificial Intelligence*, 2025. To Appear.

9 Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *CoRR*, abs/2006.11371, 2020.

10 David Debot, Pietro Barbiero, Francesco Giannini, Gabriele Ciravegna, Michelangelo Diligenti, and Giuseppe Marra. Interpretable concept-based memory reasoning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

11 Anna Himmelhuber, Stephan Grimm, Mitchell Joblin, Sonja Zillner, and Thomas A. Runkler. Combining sub-symbolic and symbolic methods for explainability. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 559–576. IOS Press, 2023.

12 Vitor A. C. Horta and Alessandra Mileo. Explaining cnns using knowledge extraction and graph analysis. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 577–608. IOS Press, 2023.

13 Filip Ilievski. *Human-Centric AI with Common Sense*. Synthesis Lectures on Computer Science (SLCS). Springer, 2024.

14 Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, Pascal Hitzler, Giuseppe Marra, Pasquale Minervini, Martin Mundt, Axel-Cyrille Ngonga Ngomo, Alessandro Oltramari, Gabriella Pasi, Zeynep G. Saribatur, Luciano Serafini, John Shawe-Taylor, Vered Shwartz, Gabriella Skitalinskaya, Clemens Stachl, Gido M. van de Ven, and Thomas Villmann. Aligning generalization between humans and machines. *Nature Machine Intelligence*, 7(9):1378–1389, Sep 2025.

15 Filip Ilievski, Kaixin Ma, Alessandro Oltramari, Peifeng Wang, and Jay Pujara. Building robust and explainable AI with commonsense knowledge graphs and neural models. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 178–209. IOS Press, 2023.

16 Youmna Ismaeil, Jan-Hendrik Metzen, Trung-Kien Tran, Hendrik Blockeel, and Daria Stepanova. Beyond manual labels: Unsupervised graph-based explanations for error analysis in image classifiers. *ISWC*, 2025. To Appear.

17 Youmna Ismaeil, Daria Stepanova, Trung-Kien Tran, and Hendrik Blockeel. Feabi: A feature selection-based framework for interpreting KG embeddings. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I*, volume 14265 of *Lecture Notes in Computer Science*, pages 599–617. Springer, 2023.

18 Youmna Ismaeil, Daria Stepanova, Trung-Kien Tran, Piyapat Saranrittichai, Csaba Domokos, and Hendrik Blockeel. Towards neural network interpretability using commonsense knowledge graphs. In Ulrike Sattler, Aidan Hogan, C. Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato, editors, *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, volume 13489 of *Lecture Notes in Computer Science*, pages 74–90. Springer, 2022.

19 Katarzyna Kaczmarek-Majer, Gabriella Casalino, Giovanna Castellano, Monika Dominiak, Olgierd Hryniewicz, Olga Kaminska, Gennaro Vessio, and Natalia Díaz Rodríguez. PLENARY: explaining black-box models in natural language through fuzzy linguistic summaries. *Inf. Sci.*, 614:374–399, 2022.

20 Xiangwei Kong, Shujie Liu, and Luhao Zhu. Toward human-centered XAI in practice: A survey. *Mach. Intell. Res.*, 21(4):740–770, 2024.

21 Q. Vera Liao, Daniel M. Gruen, and Sarah Miller. Questioning the AI: informing design practices for explainable AI user experiences. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–15. ACM, 2020.

22 Mohammad Saeid Mahdavinejad, Peyman Adibi, Amirhassan Monajemi, and Pascal Hitzler. Towards explainable depression detection: A neurosymbolic approach to uncover social media signals with generative AI. In *19th International Conference on Neurosymbolic Learning and Reasoning*, 2025.

23 Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

24 Natalia Díaz Rodríguez, Alberto Lamas, Jules Sanchez, Gianni Franchi, Ivan Donadello, Siham Tabik, David Filliat, Policarpo Cruz, Rosana Montes, and Francisco Herrera. Explainable neural-symbolic learning (*X-NeSyL*) methodology to fuse deep learning representations with expert knowledge graphs: The monumai cultural heritage use case. *Inf. Fusion*, 79:58–83, 2022.

25 Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.*, 38(5):3043–3101, 2024.

26 Timo Speith. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2239–2250. ACM, 2022.

27 Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans. Neural Networks Learn. Syst.*, 32(11):4793–4813, 2021.

28 Joe Townsend, Esma Mansouri-Benssassi, Kwun Ho Ngan, and Artur d'Avila Garcez. Discovering visual concepts and rules in convolutional neural networks. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 337–372. IOS Press, 2023.

4.5 Neurosymbolic AI in the Age of Generative AI

Daria Stepanova (Bosch Center for AI, DE), Mehwish Alam (Télécom Paris, Institut Polytechnique de Paris, FR), Stefan Ollinger (Schloss Dagstuhl – Leibniz Center for Informatics, DE)

4.5.1 What is GeNeSy AI?

Generative AI (GenAI) refers to AI systems, typically based on large-scale neural networks, that can **generate new content** such as natural language text, images, or code based on patterns learned from existing data. While the generation of **natural language text** remains the most well-known application, the scope of GenAI is much broader [47]. Typical symbolic outputs that can be produced by GenAI methods include code, formal problem specifications, structured data like graphs, tabular data, machine-readable semantic representations, logical rules and constraints (e.g., SHACL, OWL, FOL), etc. Thus, GenAI has great potential for addressing the modeling and knowledge acquisition bottleneck of symbolic AI (see Section 4.2).

4.5.1.1 GenAI Development Pipeline

A structured approach to GenAI development can be broken down into several key stages that build upon one another. It begins with data preparation, which involves sourcing, cleaning, and transforming data to create high-quality inputs for training or fine-tuning. This data may come from internal databases, third-party sources, or user-generated content. Feature engineering plays a crucial role here, helping to extract meaningful patterns while eliminating irrelevant or noisy information to ensure the model learns from the most accurate data available.

Following this, the model training phase involves developing a base generative model, either from scratch or by initializing an existing model architecture. Using large-scale datasets, the model learns to capture general patterns in multimodal data. This step typically requires extensive computing resources and represents the core of the model development process.

Once the base model is established, it is fine-tuned and customized using domain-specific or task-specific data. This tailoring process may involve supervised fine-tuning, domain adaptation, instruction tuning, or parameter-efficient techniques such as Low-Rank Adaptation, all aimed at improving the model's relevance, accuracy, and safety in real-world applications.

With a fine-tuned model, attention shifts to prompting and output control. Prompts are carefully engineered to guide the model's responses, using techniques such as zero-shot, few-shot, or chain-of-thought prompting. Sampling parameters like temperature, top-k, and

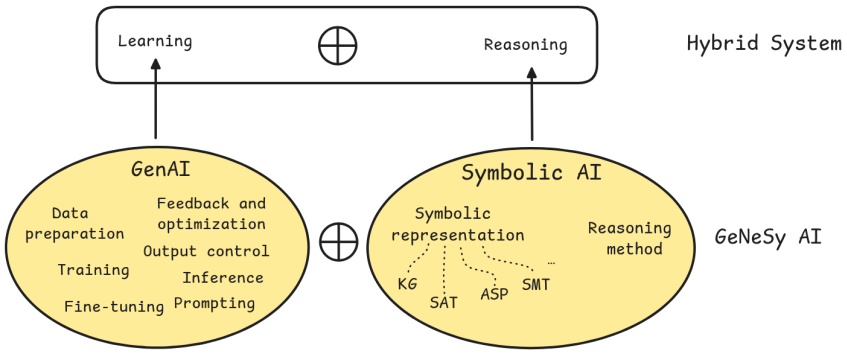


Figure 5 Illustration of a NeSy AI system, where a neural component is instantiated with the GenAI method; we refer to the resulting system as GeNeSy AI. Integration of symbolic representations, e.g., KGs, propositional formulas (SAT), answer set programs (ASP), satisfiability modular theories (SMT), etc. and the respective reasoning methods can be exploited at different stages of the GenAI development pipeline.

top-p are adjusted to manage the diversity and coherence of outputs. In the post-generation phase, outputs may undergo filtering to remove hallucinations or irrelevant content using rules, classifiers, or moderation APIs. Additional steps like re-ranking help to ensure that the most appropriate response is selected.

Finally, the feedback and optimization stage focuses on improving the model over time. Feedback is gathered through user interactions, surveys, or manual reviews to identify and address weaknesses such as bias or factual inaccuracies. This input can lead to prompt revisions, sampling tweaks, or further fine-tuning. Reinforcement learning techniques, such as Reinforcement Learning from Human Feedback (RLHF), may also be applied to better align model outputs with human preferences or task requirements.

Once development is complete, the model is deployed and integrated into applications and workflows. Post-deployment monitoring and adjustments are essential to ensure ongoing safety, reliability, and alignment with intended use.

4.5.1.2 GeNeSy AI as a Variation of NeSy AI

When it comes to NeSy AI systems in the context of GenAI, naturally the neural (learned) component of a NeSy AI system corresponds to the GenAI method itself, while the symbolic component can take various forms, such as ontologies, Knowledge Graphs (KGs), logical constraints or rules, structured schemas, or grammars (see Section 4.1). At multiple stages of the GenAI development pipeline described in Section 4.5.1.1, symbolic structures and reasoning systems can be integrated, enhancing the capabilities and reliability of generative models. We illustrate this point in Figure 1, and discuss possible integration strategies in what follows.

- **Data Preparation.** Symbolic knowledge can play a key role in generating data that conforms to predefined constraints, which can then be used–among other purposes–for training generative models. This includes, for example, the generation of graphs with specific topologies or properties, such as [30], or the synthesis of images guided by constrained scene graphs [42]. Incorporating knowledge graphs (KGs) during the data synthesis process can enhance the representativeness and structure of the generated data, e.g., see [29, 3] for question answering applications or [20] for story generation.

- **Training.** During training, symbolic structures can constrain and guide the learning process. This may involve incorporating KG embeddings, where the most prominent techniques include joint learning of Large Language Models (LLMs) and knowledge graph embeddings [9], incorporation of joint loss functions that account for both neural and symbolic objectives, embedding concatenation (see, e.g., [36] for overview), or knowledge distillation using teacher-student setups, where a symbolic model acts as a teacher [25].
- **Fine-tuning.** Symbolic knowledge, such as formal rules, specifications, or domain-specific logic, can be used to fine-tune generative models, e.g., to ensure consistency with a set of predefined rules and constraints [7]. Integration of symbolic knowledge at the fine-tuning stage is also often realized for improving performance of LLMs on structured or specialized tasks, e.g., generation of logic programs from natural language text [10]. Moreover, structural graph-based knowledge has been integrated into LLMs during fine-tuning stage to improve its performance on graph-specific tasks [38].
- **Prompting.** Symbolic knowledge can also be effectively utilized at the prompting stage, e.g., the variety of GraphRAG-based approaches are prominent examples here [14]. In the most simple scenarios factual knowledge in KGs is represented in a textual form and injected into the context during LLM prompting [5]. Prompts can also be enriched symbolically by incorporating ontological knowledge, e.g., for motion planning [11] or news summarization [43]. Other more sophisticated strategies use dedicated reasoning modules to traverse KGs and then guide the LLMs through chain-of-thought prompting [50].
- **Inference.** Finally, there are several approaches proposed in the literature that exploit symbolic reasoning to constrain or guide the output of the model during inference. Most prominent techniques include logic-based sampling [2] or decoding [40, 15, 28] to maintain consistency with the available background knowledge. Additionally, outputs produced by LLMs can be post-filtered to enforce symbolic constraints, such as exclusion of certain words or ensuring syntactic and semantic validity of generated code [27]. Several works leverage KG structures or embeddings to influence and refine the output generated by LLMs, e.g., [8]. Another promising direction is concerned with utilizing LLMs for translating textual problem statements to symbolic artifacts (see also Section 4.2), passing them to dedicated solvers for reasoning, and subsequently translating the output generated by the reasoner back to natural language using LLMs [35, 49, 12, 32].

4.5.2 What is the GeNeSy AI Ambition?

As GenAI becomes increasingly used across a variety of applications such as web search, content creation, and code generation, the goal of Neurosymbolic AI is not to replace GenAI, but to enhance it by incorporating symbolic methods, particularly in downstream tasks where GenAI tends to underperform. Recently, significant effort has been devoted to identifying and classifying such challenging tasks, as well as developing methods to address them. Much of this analysis is empirical in nature, relying on diverse benchmarks designed to highlight and characterize areas where GenAI models consistently struggle, e.g., ARC benchmarks[4]. At the same time, theoretical research has focused on approaching the problem of identifying weaknesses of GenAI methods from computational complexity and expressivity perspective.

The high-level ambition of neurosymbolic AI is that by integrating symbolic knowledge at various stages in the GenAI development pipeline as described above, GenAI systems can achieve greater accuracy, interpretability, and control, particularly in domains requiring

[4] https://arcprize.org/

structure, reasoning, or domain-specific compliance, e.g., medical applications, production systems or legal cases. Importantly, while there is a great potential for NeSy AI approaches to address many limitations of GenAI, clearly it is not a silver bullet, and precisely detecting scenarios and tasks where NeSy AI methods would have the largest impact is crucially important. Below we discuss limitations of GenAI methods, and outline possible ways how symbolic methods can be potentially utilized to address them.

Combinatorial Problems. Recently, a number of benchmarks have been introduced to evaluate the performance of GenAI methods–particularly reasoning-capable LLMs–on combinatorial tasks such as SAT solving [17], logical puzzles [26], and classical planning [21]. These studies consistently reveal a core limitation of current LLMs: their inability to handle increasing problem complexity, a phenomenon often referred to as the curse of complexity. Even with larger models and more compute, reasoning capabilities of LLMs often plateau or degrade.

While LLMs can solve certain well-known problems, such as the "25 horses" puzzle [19], their performance drops sharply on semantically equivalent but syntactically unfamiliar variants, e.g., changing "horses" to "bunnies" significantly reduces accuracy [19]. Evaluations on out-of-distribution instances [39] suggest that, in the context of combinatorial problems, many correct outputs can be attributed to memorization rather than genuine reasoning.

In contrast, symbolic AI has made considerable progress in solving these combinatorial problems effectively and accurately, as evidenced, e.g., by outcomes from SAT competitions[5]. While symbolic methods still struggle with scalability (see also Section 4.2), they offer a superior trade-off between accuracy and runtime compared to purely generative AI approaches.

This points toward the promise of neurosymbolic AI systems, which integrate LLMs with classical solvers by using LLMs to translate natural language problem descriptions into formal representations and then pass them to symbolic solvers for generating solutions [13, 32, 12, 35, 49].

Spatial and Temporal Reasoning. Recent studies have highlighted the poor performance of GenAI models on spatial and temporal reasoning tasks [4, 48, 34]. Symbolic AI, in contrast, offers mature frameworks precisely for these types of reasoning–such as RCC8 [24, 22] and Allen's interval algebra [18]. Once again, a promising direction lies in using LLMs to convert natural language inputs into formal representations that these symbolic frameworks can process [4].

Hallucinations. Hallucinations remain a critical and largely unresolved limitation of GenAI systems [6]. One potential mitigation strategy involves incorporating symbolic methods to verify or filter generated outputs. For example, [33] uses ontological reasoners to identify and eliminate hallucinations that violate logical consistency within a given ontology. Similarly, knowledge graphs have been explored as tools for grounding and validating LLM outputs [1].

Analogy and Abstraction. Human analogical reasoning involves transferring relational structures from known to novel contexts, often by applying abstract rules. This ability emerges early in development and operates across domains–from linguistic analogies (e.g., "body : feet :: table : ?") to visual ones (e.g., "(:) :: < : ?") [45]. GenAI models, however, struggle with such tasks, particularly when they involve non-standard symbols or those from low-resource languages such as Greek [45]. Analogy and abstraction has been actively studied in the area of symbolic AI, so there is a high potential for utilizing these results also in combination with GenAI.

[5] `https://satcompetition.github.io/2025/index.html`

Creative Problem Solving. The MACGYVER benchmark [46] evaluates the creative problem-solving abilities of LLMs across 1,600+ real-world scenarios that require innovative object use and out-of-the-box thinking. The results reveal that LLMs frequently suggest implausible or physically impossible actions. To address this, one promising approach involves enforcing commonsense constraints–formulated in symbolic AI languages such as first-order logic–on top of the LLM outputs.

Theoretical Limitations of Transformers. A growing body of theoretical work has identified fundamental limitations of transformer architectures, which underpin current GenAI models. For instance, it has been proven that transformers cannot natively model compositional functions [37]. As a result, they struggle with questions requiring the composition of multiple relations–such as: "What is the birthday of Frédéric Chopin's father?", which requires chaining the facts that Chopin's father is Nicolas Chopin and that Nicolas was born on April 15, 1771 [16].

These theoretical limitations have also been empirically validated in multiple studies, e.g., [23]. A promising neurosymbolic solution is to augment transformers with external knowledge representations, such as knowledge graphs [16], which naturally support compositional reasoning.

Beyond semantically rich tasks, transformers also falter on abstract algorithmic problems. For example, [41] demonstrates that even multi-head transformers cannot solve the 3-Matching problem (i.e., checking whether any three integers in a sequence sum to zero modulo a large number). While this task is synthetic, it theoretically explains the inherent limitations of transformers for combinatorial search and reasoning. In [31], the expressive power of transformers with chain-of-thought prompting is analyzed, showing that their capabilities map to the complexity class P, thus providing a theoretical upper bound on what such models can compute.

4.5.3 What are the challenges?

A central challenge in developing effective GeNeSy AI systems lies in identifying the kinds of problems where pure generative AI tends to fail, but can be meaningfully enhanced by integrating symbolic methods. Understanding these limitations is key to designing hybrid architectures that are both powerful and practical (see Section 4.8 for more details). Certain types of tasks clearly illustrate the added value of symbolic reasoning. As discussed, combinatorial problems, for instance, often go beyond the capabilities of GenAI alone and require symbolic solvers to find valid solutions or verify generated ones. Despite the combinatorial explosion of symbolic reasoning (see Section 4.2), the existing methods are still more effective than any alternatives when it comes to computing provably correct solutions.

Symbolic methods also shine in other domains, such as resolving inconsistencies across multiple sources, explaining and repairing data errors or answering complex queries with formal constraints. However, it is important to recognize that symbolic tools are not a cure-all. For example, there are failure cases, like gaps in physical world knowledge (e.g., analysis of novel diseases or structural properties of unknown materials) for which neither GenAI nor symbolic reasoning methods might be suited. The ultimate goal in this area is to develop intuitive, task-specific guidelines: for any given problem, define the GeNeSy AI configuration that is most likely to yield success.

A second major challenge lies in the emergence of suitable symbolic knowledge, both with and for GenAI. While symbolic representations like rules, constraints, or ontologies are powerful tools, they remain difficult to scale, primarily due to the bottleneck of manual

curation (see Section 4.2). A compelling direction is to explore whether GenAI itself can help construct symbolic knowledge from raw text or examples. Can GenAI be used to induce formal structures automatically? And if such structures are generated, does feeding them back into the GenAI pipeline offer measurable benefits? These are open questions that demand rigorous benchmarks to assess where and when symbolic intermediates truly improve performance.

The issue of scalability presents another critical concern. While symbolic solvers can offer precise reasoning capabilities, they often struggle to scale, especially when embedded into larger NeSy AI systems. This becomes particularly problematic when GenAI is used to generate symbolic artifacts–such as logical formulas or knowledge bases–that must then be validated or processed at scale. Tasks like solving constraints over large outputs, repairing expressive ontologies, or reasoning over extensive knowledge graphs exemplify these challenges.

In addition, aligning symbolic reasoning with multimodal GenAI systems presents unique challenges. A model processing both text and images may develop inconsistent internal representations across modalities—for instance, interpreting the concept of a "bat" as an animal in text but as a sports object in images. Such divergence can lead to errors in tasks that demand consistent cross-modal understanding. Integrating symbolic reasoning could help reconcile these discrepancies and promote coherence across modalities, but designing effective approaches remains a complex and open problem requiring careful attention.

Finally, usability of GeNeSy AI systems remains a key concern. One of GenAI's biggest strengths is its accessibility–users can interact with it naturally, without needing specialized knowledge. Symbolic systems, in contrast, often require fluency in formal languages, limiting their reach. For GeNeSy AI approaches to gain broader adoption, symbolic logic must be abstracted away from users whenever possible. The system's internal complexity should be hidden, unless deeper, expert-level access is specifically required. Ideally, interfaces with GeNeSy AI systems should be developed that retain the rigor and correctness of symbolic reasoning, while offering the ease-of-use that defines the GenAI experience. Together, these challenges frame a roadmap for building more powerful, scalable, and usable GeNeSy AI systems. They especially underscore the importance of understanding when and how to blend symbolic reasoning with GenAI to achieve the most benefit.

4.5.4 Where are we now?

A key development in addressing the limitations of generative AI lies in its growing ability to translate natural language into formal logic. This translation allows symbolic solvers, traditionally reliant on structured, logic-based input, to be effectively leveraged in tasks that originate from unstructured human language. As a result, a new wave of NeSy systems, also referred to as prompt-symbolic [44] is emerging. These systems often follow a common architecture: GenAI is used to convert natural language into precise problem specifications, which are then processed by symbolic solvers to derive solutions. In more advanced configurations, LLMs function as agents within search-based reasoning frameworks, coordinating steps in complex problem-solving tasks. This hybrid approach not only enables more rigorous reasoning but also leverages the flexibility and accessibility of GenAI to front-load formalization.

Symbolic representations are increasingly being incorporated into GenAI to enhance performance and factual accuracy. Advances such as GraphRAG illustrate this trend by integrating KGs into retrieval-augmented generation pipelines, allowing models to ground their outputs in structured knowledge. This leads to improved consistency, relevance, and logical coherence–especially useful in domains like law, where resolving contradictions across

multiple documents is critical. Moreover, symbolic methods, including the use of ontologies and KGs, support more effective integration of structured and unstructured data within GenAI systems.

Additionally, researchers are exploring closed-loop integrations of GenAI and symbolic AI, where the two components iteratively inform and refine each other. In these setups, GenAI may propose candidate solutions or knowledge structures, which are then evaluated and corrected using symbolic methods–feeding the results back into the generative process. While still an area of active exploration, such feedback loops hold promise for creating systems that not only generate and reason, but also self-correct in structured and explainable ways (see Section 4.4 for detailed discussions on explainability). Together, these developments point toward a powerful synthesis: leveraging GenAI's language capabilities to interface with symbolic tools, while embedding symbolic structure within GenAI pipelines to enhance reasoning depth and factual robustness.

References

1 Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. Can knowledge graphs reduce hallucinations in llms? : A survey. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3947–3960. Association for Computational Linguistics, 2024.

2 Kareem Ahmed, Kai-Wei Chang, and Guy Van den Broeck. Controllable generation via locally constrained resampling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

3 Fernando Amodeo, Fernando Caballero, Natalia Díaz-Rodríguez, and Luis Merino. Og-sgg: ontology-guided scene graph generation—a case study in transfer learning for telepresence robotics. *IEEE Access*, 10:132564–132583, 2022.

4 Konstantine Arkoudas. GPT-4 can't reason. *CoRR*, abs/2308.03762, 2023.

5 Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *CoRR*, abs/2306.04136, 2023.

6 Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this. *CoRR*, abs/2409.05746, 2024.

7 Diego Calanzone, Stefano Teso, and Antonio Vergari. Logically consistent language models via neuro-symbolic integration. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

8 Lang Cao. GraphReason: Enhancing reasoning capabilities of large language models through a graph-based verification approach. In Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Ben Lipkin, Danilo Neves Ribeiro, Lionel Wong, Xi Ye, and Wenting Zhao, editors, *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 1–12, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

9 Erica Coppolillo. Injecting knowledge graphs into large language models. *CoRR*, abs/2505.07554, 2025.

10 Erica Coppolillo, Francesco Calimeri, Giuseppe Manco, Simona Perri, and Francesco Ricca. LLASP: fine-tuning large language models for answer set programming. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, 2024.

11 Muhayy Ud Din, Jan Rosell, Waseem Akram, Isiah Zaplana, Máximo A. Roa, Lakmal D. Seneviratne, and Irfan Hussain. Ontology-driven prompt tuning for llm-based task and motion planning. *CoRR*, abs/2412.07493, 2024.

12 Marius-Constantin Dinu, Claudiu Leoveanu-Condrei, Markus Holzleitner, Werner Zellinger, and Sepp Hochreiter. Symbolicai: A framework for logic-based approaches combining generative models and solvers. In Vincenzo Lomonaco, Stefano Melacci, Tinne Tuytelaars, Sarath Chandar, and Razvan Pascanu, editors, *Conference on Lifelong Learning Agents, 29-1 August 2024, University of Pisa, Pisa, Italy*, volume 274 of *Proceedings of Machine Learning Research*, pages 869–914. PMLR, 2024.

13 Subhabrata Dutta, Ishan Pandey, Joykirat Singh, Sunny Manchanda, Soumen Chakrabarti, and Tanmoy Chakraborty. Frugal lms trained to invoke symbolic solvers achieve parameter-efficient arithmetic reasoning. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17951–17959. AAAI Press, 2024.

14 Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130, 2024.

15 Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10932–10952. Association for Computational Linguistics, 2023.

16 Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18126–18134. AAAI Press, 2024.

17 Rishi Hazra, Gabriele Venturato, Pedro Zuidberg Dos Martires, and Luc De Raedt. Have large language models learned to reason? A characterization via 3-sat phase transition. *CoRR*, abs/2504.03930, 2025.

18 Duygu Sezen Islakoglu and Jan-Christoph Kalo. Chronosense: Exploring temporal understanding in large language models with time intervals of events. *CoRR*, abs/2501.03040, 2025.

19 Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie Su, Camillo J. Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4722–4756. Association for Computational Linguistics, 2024.

20 Yifan Jiang, Filip Ilievski, and Kaixin Ma. Transferring procedural knowledge across commonsense tasks. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1156–1163. IOS Press, 2023.

21 Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

22 Irtaza Khalid, Amir Masoud Nourollah, and Steven Schockaert. Large language and reasoning models are shallow disjunctive reasoners. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 8843–8869. Association for Computational Linguistics, 2025.

23 Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7675–7688. Association for Computational Linguistics, 2024.

24 Fangjun Li, David C. Hogg, and Anthony G. Cohn. Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6342–6349. ijcai.org, 2024.

25 Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, and Jun Zhao. Neural-symbolic collaborative distillation: Advancing small language models for complex reasoning tasks. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 24567–24575. AAAI Press, 2025.

26 Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *CoRR*, abs/2502.01100, 2025.

27 Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. "we need structured output": Towards user-centered constraints on large language model output. In Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, and Corina Sas, editors, *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 10:1–10:9. ACM, 2024.

28 Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 780–799. Association for Computational Linguistics, 2022.

29 Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press, 2021.

30 Manuel Madeira, Clément Vignac, Dorina Thanou, and Pascal Frossard. Generative modelling of structurally constrained graphs. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

31 William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

32 Kostis Michailidis, Dimos Tsouros, and Tias Guns. Cp-bench: Evaluating large language models for constraint modelling. *CoRR*, abs/2506.06052, 2025.

33 Marco Monti, Oliver Kutz, Nicolas Troquard, and Guendalina Righetti. Improving the accuracy of black-box language models with ontologies: A preliminary roadmap. In

Proceedings of the Joint Ontology Workshops (JOWO) - Episode X: The Tukker Zomer of Ontology, and satellite events co-located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), Enschede, The Netherlands, July 15-19, 2024, volume 3882 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

34 Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. Mapwise: Evaluating vision-language models for advanced map queries. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 9348–9378. Association for Computational Linguistics, 2025.

35 Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3806–3824. Association for Computational Linguistics, 2023.

36 Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599, 2024.

37 Binghui Peng, Srini Narayanan, and Christos H. Papadimitriou. On limitations of the transformer architecture. *CoRR*, abs/2402.08164, 2024.

38 Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Seyed Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for llms. *CoRR*, abs/2402.05862, 2024.

39 Zhenting Qi, Hongyin Luo, Xuliang Huang, Zhuokai Zhao, Yibo Jiang, Xiangjun Fan, Himabindu Lakkaraju, and James R. Glass. Quantifying generalization complexity for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

40 Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 794–805. Association for Computational Linguistics, 2020.

41 Clayton Sanford, Daniel J. Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *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.

42 Giacomo Savazzi, Eugenio Lomurno, Cristian Sbrolli, Agnese Chiatti, and Matteo Matteucci. Neuro-symbolic scene graph conditioning for synthetic image dataset generation. *CoRR*, abs/2503.17224, 2025.

43 A. R. S. Silva and Y. H. P. P. Priyadarshana. Ontology-based prompt tuning for news article summarization. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2025.

44 Adam Stein, Aaditya Naik, Neelay Velingker, Mayur Naik, and Eric Wong. The road to generalizable neuro-symbolic learning should be paved with foundation models. *CoRR*, abs/2505.24874, 2025.

45 Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. Can large language models generalize analogy solving like people can? *CoRR*, abs/2411.02348, 2024.

46 Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. Macgyver: Are large language

models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5303–5324. Association for Computational Linguistics, 2024.

47 Son Tran, Edjard Mota, and Artur d'Avila Garcez. Reasoning in neurosymbolic AI. *CoRR*, abs/2505.20313, 2025.

48 Yuqing Wang and Yun Zhao. TRAM: benchmarking temporal reasoning for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6389–6415. Association for Computational Linguistics, 2024.

49 Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. Satlm: Satisfiability-aided language models using declarative prompting. 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.

50 Ruilin Zhao, Feng Zhao, Long Wang, Xianzhi Wang, and Guandong Xu. Kg-cot: Chain-of-thought prompting of large language models over knowledge graphs for knowledge-aware question answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6642–6650. ijcai.org, 2024.

4.6 Knowledge Graphs and Ontologies in Neurosymbolic Systems

Roberto Confalonieri (University of Padova, IT), Raghava Mutharaju (IIIT-Delhi, IN), Ernesto Jimenez-Ruiz (City St George's, University of London, UK), Catia Pesquita (Universidade de Lisboa, PT), Cogan Shimizu (Wright State University – Dayton, US)

4.6.1 What is it?

OWL-based knowledge graphs (KGs) [15] may play a pivotal role in NeSy systems by combining symbolic knowledge representation with formal logical reasoning. Grounded in Description Logics [2], OWL enables automated and sound reasoning for both inference (e.g., deducing new relationships) and consistency checking, offering reliable and interpretable symbolic capabilities that complement learning models. OWL-based KGs can guide learning, validate predictions, and enrich sparse or weak annotations – particularly useful in few-shot and zero-shot learning settings [8].

A major strength of OWL lies in its ability to express complex logical constraints (e.g., domain and range restrictions, class disjointness, property inverses) in a concise and reusable form. OWL-based reasoning can be embedded into neural training pipelines or used to filter semantically invalid outputs at inference time. These capabilities are supported by a rich ecosystem of open standards (e.g., [17, 7, 13]), tools (e.g., [1]), and public ontologies (e.g., [9, 3, 23]), enabling researchers to rapidly prototype and build on shared symbolic infrastructure without starting from scratch.

Reusing existing OWL ontologies and tools not only accelerates development but also promotes interoperability and reproducibility in NeSy research. Publicly available OWL-based resources provide high-quality domain knowledge and reasoning frameworks that can

be directly integrated or extended [14]. The (re)use of large ontologies and knowledge graphs to enhance neurosymbolic systems, as well as general learning and reasoning systems, is slowly gaining attention. Traditional neurosymbolic systems typically encode a limited number of rules and do not scale effectively with large modern ontologies [20].

In contrast, [5, 12] emphasize the role of OWL ontologies in generating intelligible, human-centered explanations in neurosymbolic AI. They propose a new conceptual framework highlighting three key roles of ontologies: as formal reference models, as enablers of commonsense reasoning, and as tools for abstraction and complexity management in explanations. Additionally, they introduces the idea of ontological unpacking [12] to enhance the semantic transparency of symbolic artifacts and discusses emerging challenges, such as integrating ontologies with large language models to improve trust and mitigate hallucinations.

Our NeSy position advocates for knowledge formalized through ontologies. We treat ontologies and knowledge graphs as equivalent concepts, possibly expressed in OWL or one of its fragments.[6] Our vision for NeSy falls into a more general hybrid learning and reasoning framework that (i) integrates ontologies as core components, (ii) clearly distinguishes and integrates between deductive (certain) logical reasoning and other (uncertain) learning and reasoning methods, and (iii) accommodates various forms of learning beyond neural networks. Yet, as in Section 4.1, it is also otherwise important to note that this NeSy vision does not neatly encapsulate all possible couplings between neural and symbolic components. For example, LLMs (or other neural systems capable of manipulating or interpreting text) which produce knowledge graphs or learn ontologies can be considered NeSy, as well. The nature of the composition of the components is critical, as well as their purpose.

4.6.2 Where are we now?

The neurosymbolic approach is often described in terms of a dichotomy between symbolic methods, which are human-readable and writable, and neural-based methods, which leverage connectionist training techniques [7] (see also Section 4.1). This approach employs an iterative integration cycle between symbolic and neural methods, where symbolic (expert) knowledge is embedded into neural models, and refined knowledge learned by neural networks is extracted back into symbolic form [7]. The knowledge extracted in this cycle supports the continuous refinement and modification of predefined symbolic rules through reasoning [22]. This iterative cycle is structured around three core dimensions: translation, extraction (generation), and consolidation, each emphasizing a distinct role of knowledge.

In the realm of knowledge translation, symbolic knowledge informs neural network training by embedding logical constraints directly into neural network loss functions [11] or structuring neural architectures based on established background knowledge [18]. Additionally, knowledge translation includes semantic data augmentation, which enriches training data by applying symbolic reasoning to existing datasets. This process not only improves neural network generalization [16] but also enhances the interpretability and human-understandability of subsequent explanations [6].

The **knowledge extraction** dimension focuses explicitly on deriving symbolic knowledge from trained neural networks (see also Section 4.2). Knowledge extraction is a transformation from learned representations into comprehensible rules that support explanation and reasoning tasks [21]. This extracted symbolic knowledge optimizes criteria such as accuracy, fidelity, consistency, and comprehensibility.

[6] While we choose this particular definition, other ways to define them might include that a KG is a populated ontology, or that an ontology might act as a schema for KG.

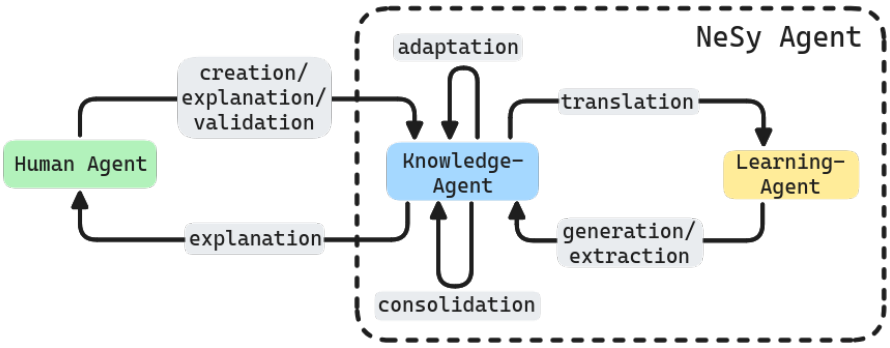


Figure 6 A proposed architecture for an ontology-mediated neurosymbolic AI.

Knowledge consolidation integrates the newly extracted symbolic knowledge back into existing knowledge structures. The revised and consolidated symbolic knowledge provides meaningful semantics to explanations, significantly facilitating human-machine interaction. Consolidated knowledge can also be reintegrated into neural models, enhancing their overall predictive performance and interpretability.

However, a more integrated approach is required, where knowledge graphs (KGs) and ontologies are treated as first-class components in neurosymbolic systems. Such integration actively involves users in the neurosymbolic cycle, empowering them to both consume and produce knowledge and explanations, thereby enabling a more robust and dynamic knowledge ecosystem.

Within this enriched integration, specific roles of knowledge emerge distinctly:

- **Knowledge creation** employs ontologies to support the formalisation of human expertise by providing a common language. This formalisation aims to capture implicit knowledge, insights, and expertise from human experts and domain specialists, structuring them explicitly using ontological languages and knowledge graphs.
- **Knowledge adaptation** employs ontology-based complexity management techniques such as abstraction, clustering, and refinement to adjust the granularity of knowledge graphs according to the demands of learning tasks and explanatory requirements.
- **Knowledge translation** leverages ontological reasoning to enrich and expand training data, guiding neural learning by explicitly embedding semantic constraints into model architectures and loss functions.
- **Knowledge generation and extraction** ensures that neural model predictions and symbolic knowledge remain consistent and interoperable, thus enhancing the reliability and transparency of derived explanations.
- **Knowledge consolidation** ensures the coherent integration of new symbolic knowledge with existing knowledge structures, maintaining consistency and semantic integrity across the entire knowledge base.
- **Knowledge explanation** grounds human-centered explanations in symbolic knowledge provided by ontologies and knowledge graphs, ensuring explanations are contextually meaningful and adapted to user expertise.

4.6.3 What is the NeSy ambition?

We envision NeSy as a hybrid agent-based architecture for continuous learning, composed of three key agents: the human agent, the knowledge-based agent, and the learning agent (Figure 6). This architecture aligns with contemporary approaches that leverage knowledge

graphs (KGs) [4] and is inherently ontology-mediated. We note that this is relatively exclusive whereby a learning agent (e.g., an LLM) is capable of *generating* or otherwise performing knowledge engineering (as in [19]). On the other hand, as these alternate methods grow in complexity (i.e., utilizing intermediate representations during the knowledge engineering tasks), it perhaps asymptotically approaches the first stated vision.

The **human agent** creates and contributes domain expertise, leveraging existing knowledge to solve specific tasks. It interacts directly with the knowledge-based agent by creating new knowledge or by **learning** new knowledge (in the form of predictions and explanations) about the system's reasoning process. Additionally, it can request further clarification if initial explanations are insufficient, actively participating in a human-in-the-loop manner to refine, consolidate and validate new knowledge.

The **knowledge-based agent** stores verified and consolidated knowledge, represented formally through ontologies and knowledge graphs. It serves the human agent by providing direct answers through symbolic inference or, when necessary, requesting new knowledge from the learning agent. Such interactions may involve **adapting** knowledge specifically for a given learning task or **translating** symbolic knowledge into semantically enriched data, or constraints into neural architectures and loss functions. The knowledge-based agent receives newly generated knowledge from the learning agent. When neural models are used, they **extract symbolic representations** in the form of structured symbolic elements, such as facts, rules, and logical statements. It **consolidates** this knowledge into existing ontologies, enriching its knowledge base while ensuring semantic consistency. This agent also delivers semantically enriched **explanations** tailored to human users.

The **learning agent** operates agnostically, capable of **generating new knowledge** either symbolically (e.g., via Inductive Logic Programming) or neurally (through deep learning methods). When neural models are used, **knowledge generation** is accompanied by the **extraction of symbolic representations** in structured symbolic forms.

In this ontology-mediated neurosymbolic framework, ontologies play pivotal roles across multiple knowledge-centric dimensions, specifically creation, adaptation, translation, generation and extraction, consolidation, and explanation.

In terms of **knowledge creation**, ontologies support the **formalization of human expertise** by providing a common language. This process captures implicit knowledge, insights, and expertise from human experts and domain specialists, structuring them explicitly using ontological languages and knowledge graphs. Ontologies facilitate knowledge creation by offering **standardized vocabularies** and **formal semantic frameworks**, enabling domain experts to express complex concepts, relationships, constraints, and rules in a precise, interoperable, and reusable manner. Such formalization ensures that previously tacit knowledge becomes explicitly available for reasoning, learning, validation, and explanation within neurosymbolic systems.

In terms of **knowledge adaptation**, ontologies enable **complexity management** through abstraction, clustering, and refinement techniques. These methods allow **fine-grained adjustments** to the granularity of the knowledge graph, aligning it closely with the specific requirements of learning tasks and the explanatory needs of end-users.

Regarding **knowledge translation**, ontologies facilitate the enrichment of training data through **semantic data augmentation**, leveraging deductive reasoning guided by ontological knowledge to enhance model inputs. Additionally, ontologies **guide feature selection** based on domain-specific semantics, thus improving the generalizability and interpretability of learning models. Furthermore, ontologies enable the **injection of logical constraints** directly into neural network training by encoding symbolic knowledge into loss functions, and they can **influence neural network architectures** by structuring them based on ontological categories and relationships, introducing effective inductive biases.

Within knowledge consolidation, ontologies facilitate the verification and validation of new knowledge generated by learning models, ensuring consistency and resolving conflicts with existing knowledge graphs. They facilitate incremental expansion of ontological frameworks, integrating newly learned symbolic knowledge in a coherent manner. Semantic interoperability is maintained by aligning new knowledge with established and standardised ontologies, promoting reusability across systems.

In terms of **knowledge generation and extraction**, ontologies help **align neural and symbolic representations**, thereby maintaining interoperability between predictions generated by neural models and the symbolic knowledge. This alignment supports the **extraction of symbolic rules and knowledge** from neural-based models, enhancing the clarity and global coherence of explanations.

Finally, for **knowledge explanation**, ontologies provide explicit **semantic grounding for explanations**, resulting in justifications that are human-understandable and contextually meaningful. Through ontology-driven abstraction, clustering, and refinement, the granularity and form of **explanations can be tailored** to the user's expertise. **Explanation-driven consolidation** further ensures transparency in knowledge integration processes by highlighting inconsistencies or unforeseen logical entailments, strengthening trust and interpretability.

4.6.4 What are the challenges, and how can we address them?

Using KGs in NeSy systems presents a range of fundamental challenges.

A crucial challenge is transforming **implicit information into explicit, machine-interpretable form**. Knowledge that is relevant to real-world tasks can potentially be inferred from data through learning or deductive reasoning, but establishing a procedure to extract and formalize this implicit knowledge, and then integrate it with established ontologies and KGs, is an open challenge (see also Section 4.2). Furthermore, properly mapping structured knowledge to raw data is crucial to ensure that the system effectively leverages the KG during both training and inference.

Another critical challenge is the **mismatch between the level of abstraction or granularity in the KG and the specific demands of the task**. KGs may contain very broad domain representations, while learning tasks often require fine-grained, context-specific features. This misalignment can degrade performance and introduce irrelevant knowledge, which adds complexity or noise. For learning agents, this may lead to suboptimal generalization or misinterpretation. For human users, it can result in confusing or overly complex explanations, undermining trust in the system's outputs. The difficulty of isolating modular, task-relevant knowledge further complicates system design and interpretability (see also Section 4.4).

Timeliness and **consistency** between the KG and real-world data also impact system performance. When new knowledge arises in the data that is not yet reflected in the KG, or when the KG evolves, but the model has not been updated. These discrepancies can lead to contradictions or inconsistencies in predictions and explanations. Effective knowledge integration mechanisms are needed to verify new information, resolve contradictions, and preserve logical consistency. **Explanations** may play a dual role here – both in surfacing inconsistencies and in making updates more interpretable to human users.

Finally, KGs pose **scalability** challenges in NeSy systems. There is often a trade-off between preserving semantic richness and interpretability, on the one hand, and achieving computational efficiency and seamless integration with neural components, on the other. High-fidelity semantic models may enable more precise reasoning and explanation, but they can be computationally intensive and harder to align with vector-based representations used

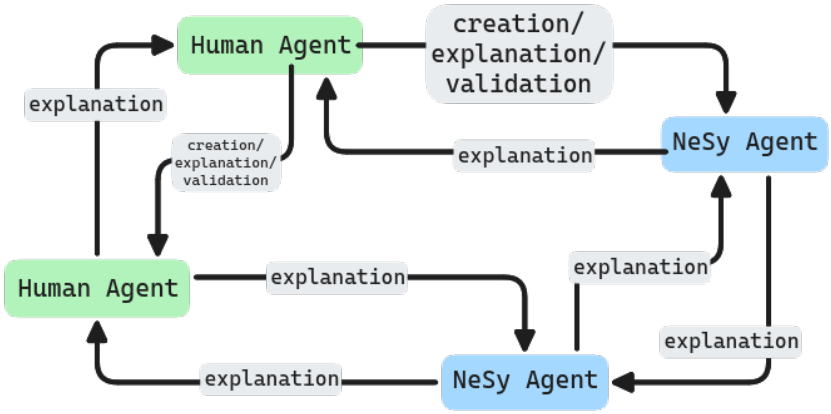


Figure 7 Multi-agent view of ontology-mediated neurosymbolic system.

in learning models. Balancing these tensions remains a central design challenge in building robust, scalable neurosymbolic AI. Figure 7 emphasizes a multi-agent view of the NeSy architecture, illustrating interactions between different types of agents (human, knowledge-based, learning), highlighting bidirectional knowledge flow and iterative refinement processes. NeSy agents can broadly include various forms, such as symbolic reasoning agents, neural agents, or even large language models (LLMs), given their symbolic output. Interaction between agents continuously updates each agent's beliefs and knowledge representations, emphasizing dynamic, evolving knowledge ecosystems (A self-loop arrow for knowledge creation might indicate self-driven learning or iterative internal refinement within human and NeSy agents).

References

1 Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

2 Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

3 BioPortal. https://bioportal.bioontology.org/.

4 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.

5 Roberto Confalonieri and Giancarlo Guizzardi. On the multiple roles of ontologies in explanations for neuro-symbolic ai. *Neurosymbolic Artificial Intelligence*, 1:NAI–240754, 2025.

6 Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296:103471, 2021.

7 Richard Cyganiak, David Wood, and Markus Lanthaler, editors. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation 25 February 2014, 2014. Available from http://www.w3.org/TR/rdf11-concepts/.

8 Claudia d'Amato, Louis Mahon, Pierre Monnin, and Giorgos Stamou. Machine learning and knowledge graphs: Existing gaps and future research challenges. *TGDK*, 1(1):8:1–8:35, 2023.

9 Dbpedia. https://wiki.dbpedia.org/.

10 Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic ai: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, Nov 2023.

11 Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. In Luc De Raedt, editor, *Proceedings of the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, IJCAI-ECAI 2022, Survey Track, Vienna, Austria, July 23-29, 2022*, pages 5478–5485. IJCAI/AAAI Press, July 2022.

12 Giancarlo Guizzardi and Nicola Guarino. Explanation, semantics, and ontology. *Data Knowl. Eng.*, 153:102325, 2024.

13 Steven Harris and Andy Seaborne. SPARQL 1.1 query language. W3C recommendation, W3C, March 2013. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

14 David Herron, Ernesto Jiménez-Ruiz, and Tillman Weyde. On the potential of logic and reasoning in neurosymbolic systems using owl-based knowledge graphs. *Neurosymbolic Artificial Intelligence*, 1:29498732251320043, 2025.

15 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021.

16 Majlinda Llugiqi, Fajar J Ekaputra, and Marta Sabou. Semantic-based data augmentation for machine learning prediction enhancement. *Neurosymbolic Artificial Intelligence*, 1:29498732251340160, 2025.

17 Boris Motik, Peter Patel-Schneider, and Bijan Parsia. OWL 2 web ontology language structural specification and functional-style syntax (second edition). W3C recommendation, W3C, December 2012. https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/.

18 Simon Odense and Artur d'Avila Garcez. A semantic framework for neurosymbolic computation. *Artif. Intell.*, 340:104273, 2025.

19 Cogan Shimizu and Pascal Hitzler. Accelerating knowledge graph and ontology engineering with large language models. *J. Web Semant.*, 85:100862, 2025.

20 Gunjan Singh, Sumit Bhatia, and Raghava Mutharaju. Neuro-symbolic RDF and description logic reasoners: The state-of-the-art and challenges. In Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 29–63. IOS Press, 2023.

21 Geoffrey G. Towell and Jude W. Shavlik. Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 13(1):71–101, Oct 1993.

22 Son Tran, Edjard Mota, and Artur d'Avila Garcez. Reasoning in neurosymbolic ai, 2025.

23 Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

4.7 Cognition and Neurosymbolic AI

Mena Leemhuis (University of Bozen-Bolzano, IT), Mehwish Alam (Télécom Paris, Institut Polytechnique de Paris, FR), Dagmar Gromann (University of Vienna, AT), Alessandro Oltramari (Bosch Center for AI, US & Carnegie Bosch Institute, US), Ute Schmid (University of Bamberg, DE), Eugene Vasserman (Kansas State University – Manhattan, US)

4.7.1 What is it?

Despite super-human performance of AI systems in very specific areas such as board games or object recognition and above average performance for tasks such as translating or summarizing texts, AI systems are lacking the flexibility, robustness, and generality of human cognition. Human cognition is based on causal models of the world allowing for understanding and explaining which is crucially different from solving pattern-recognition problems. Learning is guided by previous knowledge and experience and allows for flexible adaptation and revision based on novel information [16]. In consequence, humans can often generalize complex and productive rules from few examples [11], teach other humans about what they have learned, for instance by explanations [18], interleave learning and meta-cognition to monitor and evaluate generalizations and conclusions [32], and can apply/transfer existing knowledge to new problems and situations [17].

When human intelligence is compared with GenAI models, cognitive mechanisms such as memory, attention, and generalization are often simplified, if not completely misunderstood. For instance, the famous paper "attention is all you need", is not at all about attention, but rather focused on computational processes of working memory [31]. We tend to ascribe cognitive properties to Large Language Models (LLMs) and foundation models as a result of their human-like behavior/performance across well-defined tasks and benchmarks, although the types of errors that affect those systems are generally not observed in humans and hard to reconcile with experimental evidence from cognitive psychology. Common sense tasks, e.g., question answering, show this phenomenon quite clearly.

While LLMs show unprecedented linguistic performance, their language use is not grounded in the physical world, thus words are not connected to their real-world referents. The degree to which pattern-recognition in LLMs constitutes understanding language is debatable [7]. Human linguistic competences are formal, using the form of a language correctly, and functional, which refers to goal-directed language use [19]. To emulate both competences, models require world knowledge, ability to track changes over time, reasoning and problem-solving skills, and consideration of situative, pragmatic context [19].

Humans are often able to generalize complex and productive rules from a very small set of examples [20]. This ability is covered in many intelligence test problems such as Raven Progressive Matrices or induction of number series [10], in solving puzzles like the Tower of Hanoi [15, 28], and in generalizing relations such as "greater than" to different domains (e.g. numbers and sizes of objects). It is also apparent in language learning, for instance when learning the regular form of the past tense of verbs [20]. This cognitive ability is related to the human ability of analogy making in visual as well as in semantic domains [23]. In contrast, the data need is a severe issue of many modern AI approaches (as discussed in section "small data"). To reduce the data need, neurosymbolic architectures need to integrate perception and knowledge into generalizing from a few examples and should be based on single, generalized core mechanisms.

As discussed in section "XAI", in the context of the requirement of transparency and human control for AI systems, explainable AI (XAI) has become an active area of research [22]. Current approaches to explainability are typically one-shot and "one-size-fits-all". In contrast, explanations are central to human understanding and for the communication of causal knowledge and beliefs [18]. Although human explanations might sometimes be ex-post constructed justifications, they are considered as an important constituent of human conceptual representations [18]. Explanations in human communication are often a sequence of elucidations, varying in level of detail and modality [22, 8]. Neurosymbolic architectures should provide approaches to explainability which allow this type of context-specific adaptation to the specific information needs of a person in a given situation.

Neurosymbolic approaches have the promise to realize AI systems with more human-like abilities of learning and reasoning. AI approaches which are more closely aligned with the characteristics of human information processing may solve tasks which are better solved by humans, and also contribute to cognitive science research by providing computational models of human cognition. Finally, more human-like AI systems support better human-AI alignment for joint problem solving and decision making tasks [30].

Therefore, tackling the major goals of neurosymbolic AI, such as the data need, explainability or symbol emergence could profit from cognitively inspired approaches. This especially hints towards neurosymbolic AI, as human cognition has a tight connection to neurosymbolic ideas. In this regard, e.g., Kahneman [13] proposes with system 1 and system 2, that human cognition combines implicit and explicit reasoning. There are several viewpoints on how humans are able to model this tight integration between this implicit and explicit information. For example, Gärdenfors proposed with his Conceptual Spaces [9] one way of interpreting this human ability by modeling explicit information as geometric structures directly in the feature space.

Cognitively inspired AI approaches have a long standing tradition, e.g., prototype theory by Rosch [27] dates back to the 70's and inspired many AI approaches since then.

Other cognitively inspired AI approaches are, e.g., by [34], enhancing LLMs by incorporating attention, memory, reasoning, learning, and decision-making mechanisms or using conceptual spaces as basis for a learning approach [4]. However, they do not reach the level of tight integration needed, neither for a human-like performance nor for making up a human-machine interface.

Cognition is, however, not only of importance for reaching human-like performance due to cognitively-inspired AI. Next to that, NeSy opens up the opportunity to bridge the gap between humans and machines also in another way: by enabling human-machine collaboration by defining a neurosymbolic world model shared by human and agent.

4.7.2 What is the NeSy ambition?

Neurosymbolic AI could be used to bridge humans and machines by providing a sort of "cognitive API", thus an application programming interface for human-machine collaboration. This would allow humans and machines to negotiate a common understanding of the world ("world model") that can be expressed in a shared language. Such a world model needs to be based on a deep integration of subsymbolic and symbolic representation and reasoning: It is necessary to tightly connect explicit (maybe also shared) knowledge, e.g., in form of ontological information and implicit, feature or similarity-based information. Thus, there needs to be a grounding of the symbolic information. This especially requires an interpretation of meaning in natural language: Symbolic approaches that capture and explicitly represent world knowledge need to be integrated with powerful language representations that cater to the formal competences. This holds the promise of bringing AI systems one step closer to natural language understanding and interpretation of meaning encoded in language rather

than emulating or mimicking language behavior. This allows for both a deeper machine understanding of the human's world model and the ability of symbol emergence in the machine's world model to facilitate negotiation with a human. Construction of world models occurs by learning and generalizing from experience / data, constantly acquiring and updating knowledge.

One vital human ability needed in this context is negotiation. It is a process of disambiguation that can enable explanation, knowledge transfer (teaching, education), and collaborative decision-making. Negotiation occurs at the symbolic level, and is the iterative process used to communicate world models, achieve consensus, and make decisions.

A cognitive API should not only be able to react to symbolic negotiations but should also be able to react to non-verbal negotiations, e.g., via physical cues (especially when humans "teach" an embodied AI model how to perform physical tasks). These models must be able to "generalize", e.g., not repeat every little movement but differentiate the necessary movements from incidental ones (for instance, inferring the human intention or plan; filtering out session-specific actions as noise in multi-session training scenarios). When such inference is symbolic, we can achieve generalizable "understanding" rather than situation-specific mimicry. Physical interaction enables (semi-)autonomous control of the physical world by machine models, e.g., autonomous driving and flying. Thus, the API should be able to adapt its shared world-model in line with the human's needs.

Note that for this ambition, a cognitively inspired AI system could be beneficial, but is, however, not necessary.

4.7.3 Where are we now?

Agentic AI frameworks (see [1] for a recent survey) are used today to improve human-machine collaboration. These frameworks allow humans to execute complex tasks using LLMs connected with various software systems, such as web services, and data processing pipelines. For instance, you can ask Google Gemini to book a hotel close to where your conference is, which requires the model to use location-based services, third party websites for price comparison, reservation, etc. De facto, these systems are unidirectional: although dialog-based interaction is possible, and oftentimes necessary, there's neither an assumption nor a requirement for mutual explanation between machines and human, for bidirectional knowledge acquisition (a concept learned by a human being transferred – "taught" to a large model, and vice versa).

LLMs can simulate aspects of understanding and creativity by processing large linguistic or multimodal inputs and generating context-aware outputs. However, they often fall short in tasks involving reasoning and causal inference, which demand generalization beyond their training data [29]. To address these limitations, next to the points discussed in section "GenAI", integrating LLMs into cognitive architectures, thus systems modeled on human cognition, can enhance their robustness, adaptability, and reasoning. In this context, knowledge-grounded LLMs that use structured external knowledge are especially effective. Cognitive agents that manage reasoning, memory, or symbolic operations can further boost LLMs' capabilities and bring them closer to human-like intelligence. Still, linking LLMs to human cognition requires both theoretical and empirical validation.

Expanding cognitive architectures with multimodal inputs like eye-tracking and fMRI data can enhance alignment between artificial systems and human cognition. A benchmark dataset allowing for such examinations has recently been published [35]. Eye-tracking provides insights into attention, reading behavior, and decision-making, while fMRI captures brain activity related to language, memory, and reasoning. These allow on the one hand

to examine the alignment of AI approaches and human thought processes (as discussed, e.g., in [6]). On the other hand, integrating these signals allows cognitive models to ground language understanding in both perceptual and neural data, leading to richer, more human-like representations and responses. This approach supports the development of agents with more sophisticated, cognitively informed world models and thus communication options. However, this is a research area that is still in its infancy.

4.7.4 What are the challenges, and how to address them?

Although neurosymbolic approaches to learning and reasoning have the representational advantage of modeling both the sub-symbolic/neural and symbolic/reasoning facets of human cognition more faithfully, many of the aspects of human cognition and especially of a cognitive API discussed above are currently not or only partially addressed.

Language and Cognition. LLMs excel at mimicking human writing [7] and language. However, even in terms of formal competences, complex grammatical tasks or those that require correct semantic interpretation remain challenging, such as semantically illegal cardinality comparisons as in Fewer athletes have been to Beijing than I have [24]. Tasks that target the functional competences, including Natural Language Inference (NLI), fact checking/verification, and multi-hop question answering, are generally self-contained and represent only a small slice of the real-world. Such an approximation of meaning in AI systems might not necessarily represent understanding real-world referents [19]. A more accurate evaluation of the language-cognition alignment in AI requires more realistic benchmarks that go beyond linguistic surface forms or only small proportions of functional competences. IBM Watson's participation in Jeopardy! is a well-known example of such a challenging, knowledge-rich testbed for AI systems. A cognitive API would require such a deep real-world understanding.

Flexible re-representations for learning productive rules and abstractions from few examples. Learning complex rules and abstractions have been addressed with different, not human-like strategies [11]. For instance, generate-and-test approaches have been used to tackle the abstract reasoning challenge[7]. In contrast, humans often seem to know immediately what the relevant aspects are for generalization [14]. Purely symbolic approaches often cannot deal with noisy or imperfect input and rely on carefully tailored representations. Some autonomous driving systems, especially those that integrate world modeling / recognition and control are frequently fragile in the face of "minor" unexpected real-world artifacts, e.g., a few white dots pasted onto a stop sign or lane marker thoroughly confuses models, which may no longer see it as a traffic sign or lane marker at all; it is also possible to induce recognition as a different sign altogether [2, 3, 26, 25]. Humans are more robust in recognizing road signs, lane markers, etc., as well as inferring the plans of other drivers and pedestrians, but humans are worse at paying continuous attention to driving tasks. Furthermore, humans flexibly re-represent information in such a way that examples can be suitably aligned, that is in a goal directed manner. This has been addressed by Douglas Hofstadter with the Copycat system [12] and also by making use of background theories [33]. Current approaches to solve visual abstract reasoning problems are often neurosymbolic by combining representation learning with rule learning [21, 36]. While this seems a promising way to go, these approaches need large sets of training data and, again, re-representation is

[7] https://arcprize.org/

not addressed, restricting rule learning to the class of entities which has been present in the training examples. This especially challenges a shared human-agent world model: a shared world model is only possible when the agent is able to figure out the relevance of and can abstract from the given information.

Integrated symbolic and subsymbolic systems. Cognitive models such as Gärdenfors conceptual spaces allow for modeling a tight connection between implicit and explicit information and thus could be considered as a good strategy for modeling a tightly integrated neurosymbolic system or as a starting point for a representation of a shared world model. However, this tight connection comes to the cost of a high modeling complexity. One solution strategy is to find a trade-off between tightness of the connection between symbolic and subsymbolic and usability. Research in the area of knowledge base embeddings (KBE) [5] can be seen as a special case of conceptual spaces, where the geometric space is not aimed to model feature information but is solely focused on similarity information. This enables link prediction and query answering by modeling concept conjunction, however, it comes at the cost of losing semantic information in the form of features. This tradeoff between learnability and the representation of semantic information is a fundamental design consideration that needs to be carefully addressed when developing these cognitive inspired approaches. This also directly points towards the problem of symbol rising as discussed in the section of "symbol emergence": how can the human ability of not only reasoning with given symbols but also introducing new symbols could be handled?

Human Machine Collaboration. To conclude, all these aspects are relevant for modeling a world model shared between human and agent. However, for such a "cognitive API", there are currently not even specified requirements available. For Negotiation, what kind of language do we need? Or should we base this process on a library of languages, depending on the tasks or domains under consideration? For Construction of world models: learning is clearly a capability that both humans and machine models need, in order to distill and consolidate their knowledge. But the differences between the two modalities of learning are stark: for instance, while humans learn from few examples, neural networks require massive amounts of data. How do these different ways of learning affect the representation of the world models? Does mapping human symbols to machine-generated symbols in a common world model require mapping their different construction processes? Evaluation: how can we effectively evaluate that human-machine collaboration was successful, given a task x? Most of the current metrics focus on very specific quantitative properties of tasks (hits@k, Bleu score), however new qualitative metrics may be needed to measure collaboration. This topic is discussed in more detail in Section 4.8.

References

1 Deepak Bhaskar Acharya, Karthigeyan Kuppan, and Divya Bhaskaracharya. Agentic AI: autonomous intelligence for complex goals - A comprehensive survey. *IEEE Access*, 13:18912–18936, 2025.
2 Evan Ackerman. Three Small Stickers on Road Can Steer Tesla Autopilot Into Oncoming Lane. *IEEE Spectrum*, 2019.
3 Ross Anderson and Ilia Shumailov. Situational awareness and adversarial machine learning – robots, manners, and stress. *Apollo*, 2021.
4 Lucas Bechberger. *Using Conceptual Spaces for Artificial Intelligence*. PhD thesis, Universität Osnabrück, 2023.
5 Camille Bourgaux, Ricardo Guimarães, Raoul Koudijs, Victor Lacerda, and Ana Ozaki. Knowledge base embeddings: Semantics and theoretical properties. In Pierre Marquis,

Magdalena Ortiz, and Maurice Pagnucco, editors, *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, 2024.

6 Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), February 2022.

7 Christine Cuskley, Rebecca Woods, and Molly Flaherty. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083, 08 2024.

8 Bettina Finzel, David E. Tafler, Stephan Scheele, and Ute Schmid. Explanation as a process: User-centric construction of multi-level and multi-modal explanations. In Stefan Edelkamp, Ralf Möller, and Elmar Rueckert, editors, *KI 2021: Advances in Artificial Intelligence - 44th German Conference on AI, Virtual Event, September 27 - October 1, 2021, Proceedings*, volume 12873 of *Lecture Notes in Computer Science*, pages 80–94. Springer, 2021.

9 Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, 03 2000.

10 José V. Hernández-Conde. A case against convexity in conceptual spaces. *Synthese*, 194(10):4011–4037, Oct 2017.

11 José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L. Dowe. Computer models solving intelligence test problems: Progress and implications (extended abstract). In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 5005–5009. ijcai.org, 2017.

12 Douglas R. Hofstadter and Melanie Mitchell. The copycat project: A model of mental fluidity and analogy-making. In *Analogical connections*, Advances in connectionist and neural computation theory, Vol. 2, pages 31–112. Ablex Publishing, Westport, CT, US, 1994.

13 Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

14 Craig A Kaplan and Herbert A Simon. In search of insight. *Cognitive Psychology*, 22(3):374–419, 1990.

15 K Kotovsky, J.R Hayes, and H.A Simon. Why are some problems hard? evidence from tower of hanoi. *Cognitive Psychology*, 17(2):248–294, 1985.

16 Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *CoRR*, abs/1604.00289, 2016.

17 J. Loewenstein, L. Thompson, and D. Gentner. Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, 6(4):586–597, dec 1999.

18 Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470, 2006.

19 Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024.

20 Gary F. Marcus. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Learning, Development, and Conceptual Change. MIT Press, Cambridge, MA, 2003.

21 Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736, 2023.

22 Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

23 Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *CoRR*, abs/2102.10717, 2021.

24 Elliot Murphy, Evelina Leivada, Vittoria Dentella, Fritz Guenther, and Gary Marcus. Fundamental principles of linguistic structure are not represented by o3. *CoRR*, abs/2502.10934, 2025.

25 Ben Nassi and Yuval Elovici. Protecting autonomous cars from phantom attacks. *Communications of the ACM*, 66:56–69, March 2023.

26 Ben Nassi, Yisroel Mirsky, Dudi Nassi, Raz Ben-Netanel, Oleg Drokin, and Yuval Elovici. Phantom of the ADAS: securing advanced driver-assistance systems from split-second phantom attacks. In Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna, editors, *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 293–308. ACM, 2020.

27 Eleanor Rosch. *Principles of Categorization*, pages 312–322. Elsevier, 1988.

28 Ute Schmid and Emanuel Kitzelmann. Inductive rule learning on the knowledge level. *Cogn. Syst. Res.*, 12(3-4):237–248, 2011.

29 Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025.

30 Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, Dec 2024.

31 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

32 Marcel V. J. Veenman, Bernadette H. A. M. Van Hout-Wolters, and Peter Afflerbach. Metacognition and learning: conceptual and methodological considerations. *Metacognition and Learning*, 1(1):3–14, Apr 2006.

33 Stephan Weller and Ute Schmid. Solving proportional analogies by E -generalization. In Christian Freksa, Michael Kohlhase, and Kerstin Schill, editors, *KI 2006: Advances in Artificial Intelligence, 29th Annual German Conference on AI, KI 2006, Bremen, Germany, June 14-17, 2006, Proceedings*, volume 4314 of *Lecture Notes in Computer Science*, pages 64–75. Springer, 2006.

34 Yuanzhen Xie, Tao Xie, Mingxiong Lin, WenTao Wei, Chenglin Li, Beibei Kong, Lei Chen, Chengxiang Zhuo, Bo Hu, and Zang Li. Olagpt: Empowering llms with human-like problem-solving abilities, 2023.

35 Yunhao Zhang, Xiaohan Zhang, Chong Li, Shaonan Wang, and Chengqing Zong. Mulcog-bench: a multi-modal cognitive benchmark dataset for evaluating chinese and english computational language models. *Language Resources and Evaluation*, 59(3):3005–3028, May 2025.

36 Shukuo Zhao, Hongzhi You, Ru-Yuan Zhang, Bailu Si, Zonglei Zhen, Xiaohong Wan, and Da-Hui Wang. An interpretable neuro-symbolic model for raven's progressive matrices reasoning. *Cognitive Computation*, 15(5):1703–1724, Sep 2023.

4.8 Benchmarks in the Neurosymbolic Ecosystem

Claudia d'Amato (University of Bari, IT), Jennifer D'Souza (TIB Leibniz Information Centre for Science and Technology, DE), Anna Lisa Gentile (IBM Research Almaden, US), Hande McGinty (Kansas State University – Manhattan, US)

4.8.1 What is Benchmarking?

In computing, a benchmark is the act of running a computer program, a set of programs, or other operations, in order to assess the relative performance, normally by running a number of standard tests against it [35]. The goal of benchmarks is to provide a quantitative consensus on what constitutes good performance and to represent a shared framework for the comparison of methods [31]. Depending on the problem, there are several types of benchmarking techniques, including behavioral frameworks [34], task completion assessment [41, 3, 45], human-in-the-loop evaluations [12, 1], game-based [30, 17], etc. A dataset-based benchmark is a standardized performance test, usually consisting of a dataset or a set of datasets, a collection of questions or tasks, and a scoring mechanism including one or more metrics [31]. The task is a particular specification of the problem (as represented in the dataset). A metric is a way to summarize system performance over datasets of task(s) as a single number or score. The metric provides a means of counting success and failure at the level of individual system outputs and summarizing those counts over the full dataset [31].

4.8.1.1 Benchmark Properties, Quality and Best Practices

Various studies have attempted to provide guidelines and best practices to create effective and meaningful benchmarks, and study their desired properties. A dataset based benchmark can be formalized as a triple <dataset, task, metric>. The quality of a benchmark depends on these three components, their combination, and their usage. In [32], four main properties for minimum quality assurance have been defined, encompassing: (i) downstream utility, as grounded in real-life scenarios; (ii) validity, including size for statistical significance of results; (iii) regular updating the benchmark over time to prevent overfitting; (iv) interpretability of the score; and (iv) accessibility. A few studies propose essential guidelines to ensure good benchmarking, some addressing high level qualities such as scope or representativeness [42]; some with very detailed and granular "checklists" [7]. A well known and accepted methodology has been proposed by BetterBench, that used 46 criteria to assess the quality of a benchmark [33].

4.8.1.2 Metrics and Evaluation Methods

A metric can be defined as a computable way for measuring something quantitatively e.g, reuse metrics in software engineering, search engine quantification metrics, classification metrics, and many others. The choice of the metrics depends on the task and the properties we are interested in (for a given dataset). An example is the task of assessing the quality of software, which is typically grounded on the computation of metrics such as lines of code, cyclomatic complexity, average nesting depth, software length, effort, and time metrics [13].

In the specific perspective of neurosymbolic (NeSy) systems, multiple tasks may be of interest, including classification, semantic parsing, knowledge graph completion, visual reasoning, logical entailment, program synthesis, and symbolic planning, among others –

spanning fields such as language, vision, robotics, and ontology engineering (see Section 4.8.4 for more details). Each task can be evaluated along one or, more often, multiple metrics. For example, Classification-based metrics, such as precision, recall, F_1-score, and accuracy, are prevalent in tasks such as ontology learning, ontology alignment, semantic parsing, and visual reasoning, where discrete predictions (e.g., relation labels, answers) are evaluated against ground truth. Ranking-based metrics, such as mean reciprocal rank (MRR) and Hits@N, are standard for knowledge graph completion tasks, where models rank candidate entities. Execution-based metrics dominate in program synthesis and instruction-following benchmarks, measuring whether predicted programs (e.g., SQL queries, action plans) produce the correct output or state transition. Success rate and goal-conditioned success are often used in embodied reasoning and planning tasks, indicating whether an agent achieves the desired final state. Some benchmarks (e.g., EntailmentBank [10], ProofWriter [40]) additionally assess the quality of explanations besides the correctness of the answers, recognizing the importance of interpretable and logically coherent reasoning chains.

Beyond task-specific accuracy and ranking measures, several language generation metrics are used in translation, summarization, and structured prediction tasks. Exact match is a strict criterion, particularly valuable in question answering and program synthesis. Perplexity, originally introduced in the context of speech recognition [19], is a measure of uncertainty: the larger the perplexity, the less likely it is that an observer can guess the value which will be drawn from the distribution – in the context of LLMs it assesses how well a model predicts the next token, with lower values indicating better fluency. BLEU (Bilingual Evaluation Understudy [28]) is a measurement of the difference between an automatic translation and human-created reference translations of the same source sentence, specifically, it measures n-gram overlap in translation, while ROUGE [26], including ROUGE-N and ROUGE-L, evaluates summarization by capturing recall and sequence similarity. Together, these metrics capture complementary aspects of performance – from structural correctness and symbolic consistency to fluency and semantic alignment – highlighting the multifaceted nature of system evaluation. This diversity of metrics underscores the need for task-specific evaluation while also highlighting opportunities for unified evaluation protocols. Nevertheless, as discussed in the next section, in the specific perspective of evaluating neurosymbolic systems, there are challenges that need to be addressed, particularly aiming at assessing both neural performance and symbolic faithfulness.

4.8.2 What are the challenges, and how to address them?

Current benchmarks adopted for evaluating NeSy solutions are mostly grounded on performance with respect to defined metrics (e.g., F_1, MRR, hits@k) (see Section 4.8.4 for details). These certainly contribute to providing an objective, measurable, comparable quantification of system performances on a given task; however, they only provide a partial picture/answer to the problem. Even more so, particularly for the case of Machine Learning related tasks, evaluation revolves around inappropriate averaging/aggregation of the results [35].

For example, given two NeSy systems having rather similar F_1 values, the kind of failures may be due to different causes (e.g. errors that are due to one system being sensitive to the data distribution, errors due to the inability of the other system to capture similarities in the data space). A NeSy benchmark needs to be able to dissect the architectural components of NeSy systems in order to the assess the influence they have in the final assessment of the performances.

Similarly, given two NeSy systems having rather similar F_1 values, the current benchmarks and metrics fail to capture the influence of the neural/symbolic component respectively, on the final performance and even more so, how sensitive a system is with respect to the

(semantic) quality of each component. Indeed, by the use of symbolic (semantic) components of the NeSy systems, NeSy aims to unbox the black-box systems and inject explainability to the systems.

Another shortcoming of the current metrics and benchmarks revolve around the solutions NeSy systems can provide with small datasets. By design, NeSy systems can have the ability to provide value when datasets are small in amount or there is great internal variation within the dataset. However, current performance-based metrics may be failing to identify how a NeSy system may be able to provide better predictions based on the abstraction and symbolic components working together with numeric machine learning approaches. This kind of deeper understanding requires more fine-grained properties and metrics to be evaluated.

The analysis reported above suggests that, when talking about properties, not only benchmark properties need to be considered (see Section 4.8.1.1) but also properties referring to the (NeSy) system to be evaluated need to be taken into account, and both of them (jointly with their respective measurable metrics) should be part of the benchmark description, with a clear distinction.

Examples of benchmark properties are: dynamic benchmarking (evolving over time with respect to the dataset, keeping different dataset versions, metrics for measuring the variation within the dataset) vs. static benchmarking (stable over the time or updated from time to time) vs. real time benchmarking (benchmarks that might not exist, but need to be built in real time); data consistency vs. data inconsistency (assessed logically); benchmarks evaluating different steps along solving the targeted task (e.g. in abstract visual puzzle it is possible to distinguish the steps: perception, abstraction (what matters?), strategy; or in hypothetical reasoning different steps could be: counterfactuality, anticipatory thinking, causality) vs. benchmarks evaluating holistic solutions for the targeted task (e.g. classification).

Examples of system properties to be evaluated are: the ability to cope with/learn from small data collections vs. large data collections; the ability of explaining the learning process to the solution vs. the ability to explain/justify the solution itself disregarding the learning process vs. lack of ability of providing explanations/justifications; (lack of) robustness against data drift; verifiability, that is to be coherent with respect to existing domain knowledge; scalability and possibly also actionability, that is what can be done to change the system decision.

Last but not the least, a deeper understanding and identification of concrete problems/tasks for which the adoption of NeSy solutions in the first place is particularly beneficial is also needed.

4.8.3 What is the NeSy ambition?

Moving from the main need of having a systematic way for evaluating system/agent performances on (established) datasets by computing metrics that are relevant for assessing the system/agent ability to solve a targeted task, the main desiderata from the NeSy perspective are the ability to dissect benchmarks and evaluation under the multiple dimensions analyzed in Sect. 2, and most of all, assessing the impact and sensitivity of the characterizing (semantic) symbolic component of NeSy solutions.

Particularly, the ambition is on the definition of a generalized methodology for building benchmarks that can be operationalized and automatically customized providing the task on interest and the desired benchmark/system properties as input. While we recognize the non trivial challenge of our ambition, we evaluate it very promising both in terms of impact and feasibility. Indeed, preliminary results considering specific tasks already exist [5]. Additionally, we intend to build on existing alternative and complementary assessment

methods [31] used in different systems. These may include (but are not limited to) systematic development of test suites, audits, and adversarial testing, analyzing failure modes: system output analysis, behavioral testing, error analysis, disaggregated analysis, and counterfactual analysis, ablation testing, and analyzing model properties that are orthogonal to system outputs, such as profiling energy consumption, memory requirements, and stability in the face of perturbations to training data.

A complementary goal is the definition of a framework for checking the match/compliance of existing benchmarks with respect to (selected) properties which may enable studying/assessing the impact of these (missing) properties in solving the selected task. Additionally, assessing the fitting properties for an existing benchmark may be exploited for determining the complexity of the benchmark.

4.8.4 Where are we now?

In this section we analyze state of the art benchmarks across multiple task categories. Specifically, an overview of representative benchmarks across these task categories, including their symbolic components, domains, and evaluation metrics, is provided in Table 1. These benchmarks span a diverse set of tasks central to neurosymbolic AI, including ontology matching, knowledge graph completion, program synthesis, visual and relational reasoning, and embodied planning. Many integrate symbolic structures – such as ontologies, logical forms, scene graphs, or formal programs – with tasks that test reasoning, generalization, or action planning. Metrics range from precision/recall in classification tasks to exact match, logical form accuracy, Hits@N, and task success rate, reflecting the multifaceted nature of evaluation in this space (see also Section 4.3).

Despite this breadth, notable gaps remain. First, many benchmarks rely on static symbolic formalisms (e.g., predefined ontologies or logic rules) without testing models' ability to construct, revise, or explain symbolic representations. Second, existing datasets tend to isolate symbolic reasoning from learning under uncertainty or naturalistic conditions. Few benchmarks systematically evaluate alignment between neural and symbolic outputs, or explicitly measure interpretability, trustworthiness, or symbolic faithfulness. Moreover, coverage across scientific domains and real-world applications remains uneven – domains like biology, law, or social science are underrepresented. Addressing these limitations calls for the design of next-generation benchmarks that jointly test symbolic competence, robustness, and alignment with real-world reasoning, including multi-modal grounding.

Table 4 Benchmarks Overview.

Dataset	Task	Symbolic component	Domain	Scale	Metrics
Ontology Alignment Evaluation Initiative (OAEI) [29]	Ontology Alignment		Anatomy, Conference, Multifarm, Food, Bio-ML, Biodiversity and Ecology, Digital Humanities, Archaeology, Circular Economy, Knowledge Graphs, Pharmacogenomics		P, R, F1, Semantic P., Semantic R. [14], Runtime, consistency and conservativity [20, 37]
Large Language Models for Ontology Learning (LLMs4OL) [4, 15]	Ontology Learning		Biomedicine, Material Science, Earth and Environmental Science, Medicine, Food, Plant, Chemistry, Web		P, R, F1
FB15k-237 (from Freebase) [6]	Knowledge Graph Entity Completion	The data is essentially an ontology/graph; models often use embedding techniques but can incorporate ontological constraints or logical rules (e.g. transitivity).			Link prediction hits@N and mean reciprocal rank (MRR)
YAGO3-10	Knowledge Graph Entity Completion	As a curated ontology-derived KG, it includes type hierarchies and relational schema that methods can exploit (e.g. hasCitizenship implies a type constraint on object = Country).		123,182 entities, 37 relations, and ~1,179,040 triples, focusing heavily on persons (with relations like bornIn, hasProfession, etc.)	link prediction metrics (MRR, Hits@1/3/10)
VQA v2.0 dataset [16, 2]	Visual QA and Visual Reasoning	Questions often imply structural reasoning (counting objects, identifying attributes, etc.), though no explicit knowledge base is given.		265K images, ~1.1M questions	Acc.
Compositional Language and Elementary Visual Reasoning (CLEVR) [21]	Visual QA and Visual Reasoning	Each question comes with a functional program that specifies the reasoning steps, and ground-truth scene graphs are provided. The task is to answer complex compositional questions about the scene (counting, comparing attributes, logical operations) designed to require multi-step reasoning with minimal dataset bias.		100K images, ~865K questions	Acc.
Collision Events for Video Representation and Reasoning (CLEVRER) [43]	Video QA and Video Reasoning	Queries are designed to test understanding of physical causality; a symbolic program (event logic) can be used to derive answers		20,000 videos, annotations, and questions	Acc., per option acc., per ques acc.
Cornell Natural Language Visual Reasoning (NLVR) [38]	Visual Reasoning			92,244 pairs of examples of natural statements grounded in synthetic images with 3,962 unique sentences	Acc.
NLVR 2 [39]	Visual Reasoning			107,292 examples of English sentences paired with web photographs	Acc.
Question Answering on Image Scene Graphs (GQA Dataset) [18]	Graph QA	Images are labeled with objects, attributes, relations (scene graph); questions are represented as functional programs, enabling symbolic reasoning over the scene		113K images, 22M compositional questions	

Continued on next page

Name	Task	Symbolic component	Domain	Scale	Metrics
Outside-Knowledge VQA [27]	Visual QA and Visual Reasoning	Often leverages a knowledge base or ontology (e.g. WordNet, Wikipedia) for reasoning		~14K questions	Acc.
Spider (Text-to-SQL) [44]	Program synthesis and semantic parsing	The target output is a SQL program, and the task tests the model's ability to incorporate schema knowledge (table/column names) and logic.		10,181 natural questions and 5,693 unique SQL queries across 200 databases	Acc.
WikiSQL	Program synthesis and semantic parsing	The task is essentially mapping natural language to a SQL logical form, requiring understanding of conditions and mappings to table schema.		80,654 questions and SQL queries over 24,241 Wikipedia tables	Execution acc. (whether the predicted SQL yields the correct answer on the table) and logical form acc.
Compositional Freebase Questions (CFQ) [23]	semantic parsing	Uses a fixed ontology (Freebase schema) and logical queries (SPARQL)		240k natural questions generated from Freebase (with answers), each paired with a SPARQL query against a knowledge graph	Acc. and compound divergence metric.
SCAN (Simplified versions of the CommAI Navigation tasks) [24]	semantic parsing	The action sequence can be seen as a program the agent must generate, and generalization requires following combinatorial rules (e.g. learning the meaning of "twice").			Exact match of the output action sequence; various splits test zero-shot compositional generalization
ARC (Abstraction and Reasoning Corpus) [9]	semantic parsing	The underlying solution for each puzzle is effectively a small program or rule (e.g. "reflect the pattern", "count and place objects") that the AI must infer. No training examples are provided per task, emphasizing few-shot abstract reasoning		contains 1,000 grid-based puzzles where a small set of input-output examples is given and the model must output the result for a new input	Number of tasks solved perfectly (traditionally, execution on hidden test cases)
The "Compositional Language Understanding and Text-based Relational Reasoning" benchmark (CLUTRR) [36]	Logical and Relational Reasoning	Under the hood, each story corresponds to a small knowledge graph of family relations and a logical proof chain; the dataset systematically varies the number of hops and adds distracting facts to test inductive reasoning robustness			Relation prediction acc.
Higher-order Logic Theorem Proving (HolStep) [22]	Logical and Relational Reasoning	Each example is a formal logic formula or theorem; solving tasks involves logical inference in a strict symbolic sense (neural models must interface with formal rules).		2 million logical statements and 10,000 theorems from higher-order logic proofs in the HOL Light system	Usually reported as precision/recall for premise selection or accuracy of proof step prediction.
TPTP library ("Thousands of Problems for Theorem Provers")	Logical and Relational Reasoning	All problems are given in formal logic syntax (CNF or TPTP format); automated theorem provers (ATPs) or neuro-symbolic reasoners attempt to prove a conjecture from given axioms.			number of problems solved and proof quality

Continued on next page

Name	Task	Symbolic component	Domain	Scale	Metrics
EntailmentBank [11] & ProofWriter [40]	Logical and Relational Reasoning	The provided explanations are structured as logical entailment proofs (often in natural language form, but representing a symbolic proof structure) which models are encouraged to reproduce. These benchmarks push models to perform symbolic reasoning (like chaining facts) in addition to answering correctly.			twofold – accuracy of the final answer and some measure of explanation quality
BabyAI (instruction-following in a gridworld) [8]	Robotics and Planning	The environment is modeled as a grid with objects; an instruction corresponds to a structured plan (sequence of actions) that could be represented in a formal language. The BabyAI platform even includes a built-in verifier that symbolically checks if an agent's actions satisfy the command		There are 19 BabyAI levels with increasing complexity, testing sequence of skills and compositional learning.	Success rate on completing the instruction correctly, as well as sample efficiency (learning from few demonstrations).
ALFRED (Action Learning From Realistic Environments and Directives) [34]	Robotics and Planning	Tasks have an underlying plan structure (open fridge -> grab apple -> heat apple -> etc.); they can be represented as sequences of high-level actions. ALFRED annotations include step-by-step action sequences as the ground-truth plan.		1,000+ household tasks described by natural language instructions and visual observations	Success rate and goal-condition success (did the final world state meet the objective). This tests integration of vision (for perception) with symbolic planning/execution.
Neuro-Symbolic VQA/NS-CL [43]	Robotics and Planning	The model uses an explicit program (in a logical DSL) parsed from language, which is then executed on a neural representation of the scene			Measures include question-answering accuracy and generalization to new combinations of attributes or novel tasks.
Neuro-Symbolic Action Planning (NS-AP)	Robotics and Planning	Tasks are described as sequences of subgoals that can be represented by a symbolic program, analogous to a subroutine in a classical planner			Success rate on 10 benchmarking scenarios, which require correctly executing all subgoals.
LogiCity – A simulated urban environment benchmark for neuro-symbolic reasoning in dynamic scenes [25]	Robotics and Planning	The environment's dynamics are defined by FOL rules (e.g. logic clauses for when an entity must stop or yield)	It defines an urban world with cars, pedestrians, etc., and customizable first-order logic rules governing their behavior (e.g. traffic rules, right-of-way)		In navigation, success rate of reaching the goal safely; in action prediction, accuracy of predicting correct agent actions per the logic.

References

1 Maryam Amirizaniani, Jianli Yao, Andrew Lavergne, Edward S. Okada, Aman Chadha, Tina Roosta, and Chirag Shah. Llmauditor: A framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*, 2024.

2 Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

3 Negar Arabzadeh, Siqing Huo, Nikhil Mehta, Qingyun Wu, Chi Wang, Ahmed Awadallah, Charles L.A. Clarke, and Julia Kiseleva. Assessing and verifying task utility in llm-powered applications. *arXiv preprint arXiv:2405.02178*, 2024.

4 Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol: Large language models for ontology learning. In *International Semantic Web Conference*, pages 408–427. Springer, 2023.

5 Roberta Barile, Claudia d'Amato, and Nicola Fanizzi. Lp-dixit: evaluating explanations for link predictions on knowledge graphs using large language models. In *Proceedings of the ACM on Web Conference 2025 (WWW)*, WWW '25, pages 4034–4042, Sydney, NSW, Australia, 2025. ACM.

6 Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

7 Junjie Cao, Yiu-Kwan Chan, Zhihua Ling, Wei Wang, Shaowu Li, Ming Liu, Rong Qiao, Yunlong Han, Chenglei Wang, Bowei Yu, Peng He, Shilin Wang, Zhenming Zheng, Michael R. Lyu, and S.C. Cheung. How should we build a benchmark? revisiting 274 code-related benchmarks for llms. *arXiv preprint arXiv:2501.10711*, 2025.

8 Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Sacha Lahlou, Lucas Willems, Chimera Saharia, Thang H. Nguyen, and Yoshua Bengio. BabyAI: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272*, 2019.

9 François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

10 Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. *EMNLP*, 2021.

11 Bhushan Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnie Xie, Hannah Smith, Lalit Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. *arXiv preprint arXiv:2104.08661*, 2022.

12 I. Drori and D. Te'eni. Human-in-the-loop ai reviewing: feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109, 2024.

13 H.E. Dunsmore. Software metrics: An overview of an evolving methodology. Technical report, Technical Report, 1984.

14 Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 348–353, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

15 Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol 2024 overview: The 1st large language models for ontology learning challenge. In *Open Conference Proceedings*, volume 4, pages 3–16, 2024.

16 Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

17 Lichuan Hu, Qian Li, An Xie, Ning Jiang, Ion Stoica, Hai Jin, and Haichao Zhang. Gamearena: Evaluating llm reasoning through live computer games. *arXiv preprint arXiv:2412.06394*, 2024.

18 Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

19 F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity – a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005.

20 Ernesto Jiménez-Ruiz, Christian Meilicke, Bernardo Cuenca Grau, and Ian Horrocks. Evaluating mapping repair systems with large biomedical ontologies. In Thomas Eiter, Birte Glimm, Yevgeny Kazakov, and Markus Krötzsch, editors, *Informal Proceedings of the 26th International Workshop on Description Logics, Ulm, Germany, July 23 – 26, 2013*, volume 1014 of *CEUR Workshop Proceedings*, pages 246–257. CEUR-WS.org, 2013.

21 Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

22 Cezary Kaliszyk, François Chollet, and Christian Szegedy. HOLStep: A machine learning dataset for higher-order logic theorem proving. *arXiv preprint arXiv:1703.00426*, 2017.

23 Daniel Keysers, Nathanael Schärli, Nathan Scales, H. Buisman, D. Furrer, S. Kashubin, N. Momchev, D. Sinopalnikov, L. Stafiniak, and T. Tihon et al. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations (ICLR)*, 2020.

24 Brendan M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*, 2018.

25 Bowen Li, Zhaoyu Li, Qiwei Du, Jinqi Luo, Wenshan Wang, Yaqi Xie, Simon Stepputtis, Chen Wang, Katia P. Sycara, Pradeep Kumar Ravikumar, Alexander Gray, Xujie Si, and Sebastian Scherer. LogiCity: Advancing neuro-symbolic ai with abstract urban simulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

26 Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, July 2004.

27 Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, 2022.

28 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

29 M. A. N. Pour, A. Algergawy, E. Blomqvist, P. Buche, J. Chen, P. G. Cotovio, A. Coulet, J. Cufi, H. Dong, D. Faria, L. Ferraz, S. Hertling, Y. He, I. Horrocks, L. Ibanescu, S. Jain, E. Jiménez-Ruiz, N. Karam, F. Kraus, P. Lambrix, H. Li, Y. Li, P. Monnin, H. Paulheim, C. Pesquita, A. Sharma, P. Shvaiko, M. Silva, G. Sousa, C. Trojahn, J. Vataščinová, B. Yaman, O. Zamazal, and L. Zhou. Results of the ontology alignment evaluation initiative 2024. In *19th International Workshop on Ontology Matching*, volume 3897, pages 64–97, January 2025. © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

30 Deyi Qiao, Chengming Wu, Yilun Liang, Junyi Li, and Nan Duan. Gameeval: Evaluating llms on conversational games. *arXiv preprint arXiv:2308.10032*, 2023.

31 Inioluwa Deborah Raji, Emily M. Denton, Emilly M. Bender, Amandalynne Paullada, and A. T. Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of*

the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS), volume 1, 2021.

32 Andrew Reuel, Anthony Hardy, Carson Smith, Matthew Lamparth, Michelle Hardy, and Mykel J. Kochenderfer. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. *arXiv preprint arXiv:2411.12990*, 2024.

33 Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912. Association for Computational Linguistics, 2020.

34 Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yejin Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *arXiv preprint arXiv:1912.01734*, 2020.

35 Edgar H. Sibley, Philip J. Fleming, and John J. Wallace. Computing practices how not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM*, 29:218–221, March 1986.

36 Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*, 2019.

37 Alessandro Solimando, Ernesto Jiménez-Ruiz, and Giovanna Guerrini. Minimizing conservativity violations in ontology alignments: algorithms and evaluation. *Knowl. Inf. Syst.*, 51(3):775–819, 2017.

38 Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July 2017. Association for Computational Linguistics.

39 Alane Suhr, Siyuan Zhou, Angli Zhang, Iris Zhang, Huajie Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019.

40 Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online, August 2021. Association for Computational Linguistics.

41 Ziyu Wang, Siyuan Zhao, Yutong Wang, Hao Huang, Sicheng Xie, Yi Zhang, Jiashi Shi, Zehua Wang, Hai Li, and Jianjun Yan. Re-task: Revisiting llm tasks from capability, skill, and knowledge perspectives. *arXiv preprint arXiv:2408.06904*, 2024.

42 Lukas M. Weber, Wouter Saelens, Robrecht Cannoodt, Charlotte Soneson, Alexander Hapfelmeier, Paul P. Gardner, Anne-Laure Boulesteix, Yvan Saeys, and Mark D. Robinson. Essential guidelines for computational method benchmarking. *Genome Biology*, 20(125), 6 2019.

43 Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural- symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2019.

44 Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, 2018.

45 Rui Zhou, Liang Chen, and Ke Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024.

4.9 Real-World Applications in Neurosymbolic Artificial Intelligence

Ernesto Jimenez-Ruiz (City St George's, University of London, UK), Roberto Confalonieri (University of Padova, IT), Mena Leemhuis (University of Bozen-Bolzano, IT), Catia Pesquita (Universidade de Lisboa, PT), Daria Stepanova (Bosch Center for AI, DE)

4.9.1 What is it?

Over the past decade, AI development has largely focused on data-driven, output-oriented models. More recently, however, there has been a notable shift toward approaches that prioritize trust, transparency, and control. In response to this shift, neurosymbolic AI has gained increasing attention, as it integrates symbolic reasoning with data-driven methods to support more interpretable and reliable systems. As a result, a growing number of neurosymbolic approaches are now being explored and adopted in practice.

4.9.1.1 Key Domains of Application

NeSy AI is gaining traction across a wide range of real-world application domains, particularly in areas where interpretability, explainability, and trust are critical.

In **scientific discovery**, for instance, drug development benefits from systems like IBM's RXN[8], which pair neural networks for reaction prediction with symbolic logic to ensure chemical validity. Similarly, physics research leverages NeSy AI models to derive symbolic equations from experimental data, aiding in interpretable and data-driven discovery. Recently, also startups, e.g., extensity[9] started developing neurosymbolic platforms to accelerate scientific research work by integrating LLMs and knowledge graph technologies.

In **healthcare and medical diagnostics**, radiology applications combine image-based neural diagnosis with symbolic clinical guidelines to support interpretable decision-making, e.g., in [9] this approach has been applied to the problem of Alzheimer's disease diagnostics. Another application is concerned with infection tracking and patient flow, where data about patients and staff movement (collected via tracking devices) is combined with hospital constraints and domain knowledge represented as clinical rules (capturing infection policies) and knowledge graphs (hospital roles, wards, rooms, points of interest, etc.) [14]. In [13] Logical Neural Networks (LNNs) are employed to embed domain-specific rules as weighted logical formulas for explainable diagnosis. Additionally, [7] demonstrates how symbolic reasoning enhances automated radiology report generation by grounding neural outputs in structured clinical logic.

In **industrial automation**, quality control benefits from neural image inspection systems supported by symbolic rule enforcement to ensure product standards. For example, in [1] a neuro symbolic AI approach for automating the compliance verification of the electrical control panels has been proposed. This method combines Deep Learning techniques for recognizing the electrical components from the images of the electrical control panels with an Answer Set Programming-based system for comparing the scheme reconstructed from the picture with its original version to discover possible errors. The work [5] focuses on

[8] https://rxn.app.accelerate.science/
[9] https://www.extensity.ai/

the problem of predictive maintenance, where a neurosymbolic architecture is designed to both detect anomalies and explain their causes. It combines a state-of-the-art unsupervised autoencoder for anomaly detection with an online rule-learning algorithm that provides symbolic explanations for anomalies.

Neurosymbolic AI methods are also applied for **system configuration problems**, e.g., for configuration of E-Drives, where answer set programming (ASP) based approaches are used for finding conceptual designs of E-Drives that satisfy user requirements, and LLMs are invoked to facilitate the interaction with the ASP solver by translating formal explanations to natural language [6]. In a similar way, there is a tendency of combining more classical approaches for scheduling and optimization, e.g., [3, 4] with LLMs to facilitate the interaction with the system. Other NeSy AI applications include, e.g., welding quality monitoring [19] or root cause analysis [17].

In **supply chain optimization**, traditional predictive models often fail to consider complex legal, contractual, and sustainability constraints, leading to infeasible or non-compliant recommendations. Neurosymbolic AI offers a solution by integrating neural forecasting models with symbolic rule-based systems that enforce supplier agreements, trade compliance laws, and environmental policies. These systems can generate optimized yet compliant strategies across global supply chains. For example, knowledge graph embeddings have been applied for retrieving supplier candidates that are similar to the currently available suppliers, and ASP has been used for finding the most optimal selection of suppliers which minimize a certain objective, e.g., CO2 emission [18]. Moreover, product intelligence and trusted traceability tools have been developed that leverage symbolic logic to ensure regulatory compliance while optimizing logistics operations [16].

Other growing domains of neurosymbolic AI applications include **finance**, **e-commerce**, and **cybersecurity**, where hybrid models help to enforce regulations, detect anomalies, and explain decisions, which are critical capabilities in high-risk and fast-evolving environments. Traditional fraud detection systems often rely on historical transaction patterns, which can quickly become outdated as fraud tactics evolve. Neurosymbolic AI addresses this limitation by combining neural anomaly detection, which can scan millions of transactions for suspicious behavior, with symbolic reasoning that enforces logical constraints, such as regulatory anti-money laundering (AML) rules. This hybrid approach flags potentially fraudulent activity and provides interpretable explanations of which compliance rules were violated, increasing transparency for investigators and auditors. The work [2] integrates a transformer-based neural model with a symbolic Belief-Desire-Intention (BDI) reasoning layer, thus significantly improving the interpretability and decision-making capabilities of fraud detection pipelines. Recently, several startups have emerged that provide hybrid solutions for banking and financial applications, e.g., chatbots that comply with audit-based regulations[10].

Robotics and autonomous systems benefit from neurosymbolic planning, where neural perception handles sensory input, and symbolic reasoning governs task execution and long-term strategy. In human-robot interaction, language understanding (via neural networks) is paired with symbolic logic for command execution.

In the **legal domain**, contract analysis systems use neural NLP models to process text and symbolic reasoning to validate legal clauses. Legal compliance platforms also combine machine learning with rule-based systems to ensure business processes align with regulations. Emerging research even explores how large language models (LLMs) can be evaluated for legal soundness by linking their outputs with specific legal texts through symbolic frameworks

[10] `unlikely.ai`

and prompting techniques. Some works analyze the legal implication of answers provided by LLMs and possibly alert the user. Prompting techniques and RAG are mostly adopted while actual references to the interested law and corresponding articles can be provided by suitable KGs [8]

In **education**, intelligent tutoring systems use neural models to understand inputs like handwriting or speech and symbolic solvers to guide students through structured reasoning in subjects like mathematics. Adaptive learning platforms track student behavior and apply logic-based strategies to personalize learning paths [11].

In **games and strategy applications**, Neurosymbolic AI combines neural network learning with symbolic reasoning to enable dynamic planning and strategy formulation, as seen in complex gaming environments (for example, SwarmBrain in StarCraft II, which uses LLMs for macro-strategy and symbolic or rule-like control modules for tactical execution) [15].

4.9.2 What is the NeSy ambition?

Neurosymbolic AI aims to address several core limitations of purely neural systems. First and foremost, it enhances **interpretability and explainability** (see Section 4.4), particularly crucial in domains like healthcare, finance, or law, where opaque models are often unsuitable due to safety, ethical, or regulatory concerns. By incorporating symbolic reasoning, systems become more transparent and human-compatible, enabling more robust human-in-the-loop applications.

Another key advantage is **better generalization from small datasets** (see Section 4.3). Traditional deep learning often requires large volumes of data, which is not always available in practice due to rarity, privacy concerns, or high acquisition costs. NeSy AI models can bridge data gaps by incorporating domain knowledge through symbolic representations, allowing for meaningful inferences even with limited data.

Regulatory compliance and ethical alignment are also central to the promise of NeSy AI. In domains governed by strict rules and societal expectations, the ability to directly encode laws, guidelines, or ethical norms into the symbolic component of a model ensures more predictable and auditable behavior.

From a development perspective, **debugging and maintenance** are easier in neurosymbolic systems compared to opaque neural models. The symbolic component offers points of inspection and control, simplifying the identification and resolution of issues.

NeSy AI systems also show promise in **handling out-of-distribution inputs**. Real-world environments are unpredictable, and systems must handle novel scenarios. Symbolic rules–such as safety constraints or physical laws–act as safeguards, enabling the system to maintain functionality even in unfamiliar situations. Interestingly, symbolic knowledge doesn't just constrain outputs but can also expand them, guiding the model toward new, valid solutions not directly observed in training data.

Additionally, in many real-world applications there is a growing need to enable effective collaboration between humans and neural models–particularly large language models (LLMs)–in fields such as systems engineering and production optimization [12]. Addressing this challenge requires hybrid, agent-based architectures capable of supporting continuous learning, reasoning, and knowledge exchange. The architectures proposed in Section 4.6, which incorporate not only neural and human agents but also a knowledge-based agent–aim to fulfill this need by facilitating meaningful interaction between humans and machines. This enables effective and transparent knowledge sharing and decision support. Unlike purely neural approaches (e.g., LLM-based systems) or purely symbolic systems (e.g., classical expert systems), these hybrid architectures combine learning and reasoning capabilities, thereby enabling a continuous cycle of adaptation and improvement.

4.9.3 What are the challenges?

Despite their promise, NeSy AI approaches face several notable challenges. One of the main ones among them is the **lack of a clear methodology** for designing and building such systems (see also Section 4.1). Transforming raw data or implicit expert knowledge into symbolic representations is often labor-intensive and requires deep domain expertise (see also Section 4.2).

Many current NeSy AI implementations are still limited to small datasets or artificial benchmarks, lacking demonstration of **scalability and real-world applicability**. Real-world environments change rapidly, raising questions about how to maintain **synchronization between evolving knowledge models and learning systems**.

Moreover, while symbolic reasoning enhances explainability, it can sometimes reduce the raw performance or efficiency of systems, especially when fast inference is critical–as in robotics, finance, or emergency medicine. Symbolic components can also add computational overhead, particularly when combined with large-scale neural models (see again Section 4.2).

There are also practical limitations regarding **tooling and infrastructure**. Unlike the mature ecosystems for deep learning (e.g., PyTorch, TensorFlow), few production-ready libraries support neurosymbolic development. This has recently started to change, however, as new startups focusing on neurosymbolic integrations start developing their own advanced platforms, e.g., Imandra[11]. Still evaluating these systems at large scale in real-world settings remains difficult, with no standard benchmarks and challenges around integrating human factors like usability and interface design (see Section 4.8).

4.9.4 Where are we now?

A foundational requirement for neurosymbolic AI is access to structured knowledge–ontologies, knowledge graphs, argumentation structures–that serve as the backbone of the symbolic component. Historically, the scarcity of such data has limited NeSy AI systems in real-world contexts (see Section 4.2).

However, recent advances in large language models have transformed this landscape. Research efforts, such as those by Hu et al. [10], demonstrate how LLMs can be used to automatically generate structured data, including knowledge graphs and ontologies. This unlocks neurosymbolic applications even in niche domains where no prior structured knowledge existed.

While this marks significant progress, it introduces new concerns around **data quality, provenance, and correctness**. Human-in-the-loop mechanisms become essential to validate and refine LLM-generated knowledge, ensuring the symbolic backbone of NeSy AI systems remains reliable. Nonetheless, these developments substantially broaden the scope of neurosymbolic AI, making its core benefits–interpretability, trustworthiness, and data efficiency–more accessible than ever before.

Neurosymbolic AI represents a promising paradigm shift in the development of intelligent systems. By integrating neural and symbolic approaches, it offers a path toward AI that is not only powerful but also transparent, adaptable, and aligned with human values and regulatory frameworks. While the field still faces methodological and practical hurdles, the rapid progress in LLM-driven knowledge generation and growing interest across domains signal a strong trajectory forward. As the ecosystem matures, NeSy AI approaches may become foundational in building AI we can truly understand and trust.

[11] `https://imandra.ai/`

References

1. Vito Barbara, Massimo Guarascio, Nicola Leone, Giuseppe Manco, Alessandro Quarta, Francesco Ricca, and Ettore Ritacco. Neuro-symbolic AI for compliance checking of electrical control panels. *Theory Pract. Log. Program.*, 23(4):748–764, 2023.
2. Parul Dubey, Pushkar Dubey, and Pitshou N. Bokoro. A unified transformer–bdi architecture for financial fraud detection: Distributed knowledge transfer across diverse datasets. *Forecasting*, 7(2), 2025.
3. Thomas Eiter, Tobias Geibinger, Nysret Musliu, Johannes Oetsch, Peter Skocovský, and Daria Stepanova. Answer-set programming for lexicographical makespan optimisation in parallel machine scheduling. *Theory Pract. Log. Program.*, 23(6):1281–1306, 2023.
4. Thomas Eiter, Tobias Geibinger, Nelson Higuera Ruiz, Nysret Musliu, Johannes Oetsch, Dave Pfliegler, and Daria Stepanova. Adaptive large-neighbourhood search for optimisation in answer-set programming. *Artif. Intell.*, 337:104230, 2024.
5. João Gama, Rita P. Ribeiro, Saulo Martiello Mastelini, Narjes Davari, and Bruno Veloso. A neuro-symbolic explainer for rare events: A case study on predictive maintenance. *CoRR*, abs/2404.14455, 2024.
6. Tobias Geibinger, Tobias Kaminski, and Johannes Oetsch. Explanations for guess-and-check asp encodings using an llm (extended abstract). In *Program TAASP. Workshop on Trends and Applications of Answer Set Programming (TAASP 2024)*, page 6, Klagenfurt, Austria, 2024.
7. Zhongyi Han, Benzheng Wei, Xiaoming Xi, Bo Chen, Yilong Yin, and Shuo Li. Unifying neural learning and symbolic reasoning for spinal medical report generation. *Medical Image Anal.*, 67:101872, 2021.
8. George Hannah, Rita T. Sousa, Ioannis Dasoulas, and Claudia d'Amato. On the legal implications of large language model answers: A prompt engineering approach and a view beyond by exploiting knowledge graphs. *Journal of Web Semantics*, 84:100843, 2025.
9. Yexiao He, Ziyao Wang, Yuning Zhang, Tingting Dan, Tianlong Chen, Guorong Wu, and Ang Li. Neurosymad: A neuro-symbolic framework for interpretable alzheimer's disease diagnosis. *CoRR*, abs/2503.00510, 2025.
10. Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. Enabling llm knowledge analysis via extensive materialization, 2025.
11. Chris Davis Jaldi, Eleni Ilkou, Noah L. Schroeder, and Cogan Shimizu. Education in the era of neurosymbolic AI. *J. Web Semant.*, 85:100857, 2025.
12. Sebastian Krakowski. Human-ai agency in the age of generative ai. *Information and Organization*, 35(1):100560, 2025.
13. Qiuhao Lu, Rui Li, Elham Sagheb, Andrew Wen, Jinlian Wang, Liwei Wang, Jungwei W. Fan, and Hongfang Liu. Explainable diagnosis prediction through neuro-symbolic integration. *CoRR*, abs/2410.01855, 2024.
14. Chi Him Ng, Annette ten Teije, and Frank van Harmelen. A boxology-based analysis of design patterns for neuro-symbolic medical decision making systems. In Riccardo Bellazzi, José Manuel Juarez Herrero, Lucia Sacchi, and Blaz Zupan, editors, *Artificial Intelligence in Medicine – 23rd International Conference, AIME 2025, Pavia, Italy, June 23-26, 2025, Proceedings, Part I*, volume 15734 of *Lecture Notes in Computer Science*, pages 333–343. Springer, 2025.
15. Xiao Shao, Weifu Jiang, Fei Zuo, and Mengqing Liu. Swarmbrain: Embodied agent for real-time strategy game starcraft ii via large language models, 2024.
16. Siemens Digital Industries Software. Product intelligence & trusted traceability, 2023.
17. Nicholas Tagliapietra, Juergen Luettin, Lavdim Halilaj, Moritz Willig, Tim Pychynski, and Kristian Kersting. Causalman: A physics-based simulator for large-scale causality. *CoRR*, abs/2502.12707, 2025.

18 Cuong Chu Xuan, Mohamed H. Gad-Elrab, Trung-Kien Tran, Marvin Schiller, Evgeny Kharlamov, and Daria Stepanova. Supplier optimization at bosch with knowledge graphs and answer set programming. In *The Semantic Web: ESWC 2023 Satellite Events – Hersonissos, Crete, Greece, May 28 – June 1, 2023, Proceedings*, volume 13998 of *Lecture Notes in Computer Science*, pages 200–204. Springer, 2023.

19 Baifan Zhou, Zhipeng Tan, Zhuoxun Zheng, Dongzhuoran Zhou, Yunjie He, Yuqicheng Zhu, Muhammad Yahya, Trung-Kien Tran, Daria Stepanova, Mohamed H. Gad-Elrab, and Evgeny Kharlamov. Neuro-symbolic AI at bosch: Data foundation, insights, and deployment. In Anastasia Dimou, Armin Haller, Anna Lisa Gentile, and Petar Ristoski, editors, *Proceedings of the ISWC 2022 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 21st International Semantic Web Conference (ISWC 2022), Virtual Conference, Hangzhou, China, October 23-27, 2022*, volume 3254 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

Participants

- Mehwish Alam
Institut Polytechnique de Paris, FR
- Vaishak Belle
University of Edinburgh, GB
- Roberto Confalonieri
University of Padova, IT
- Claudia d'Amato
University of Bari, IT
- Artur d'Avila Garcez
City – University of London, GB
- Jennifer D'Souza
Leibniz Universität Hannover, DE
- Luc De Raedt
KU Leuven, BE
- Natalia Díaz-Rodríguez
University of Granada, DE
- Anna Lisa Gentile
IBM Almaden Center – San Jose, US
- Dagmar Gromann
Universität Wien, AT
- Pascal Hitzler
Kansas State University – Manhattan, US
- Filip Ilievski
VU Amsterdam, NL
- Ernesto Jiménez-Ruiz
City – University of London, GB
- Mena Leemhuis
Free University of Bozen-Bolzano, IT
- Bertram Ludäscher
University of Illinois at Urbana-Champaign, US
- Giuseppe Marra
KU Leuven, BE
- Hande McGinty
Kansas State University – Manhattan, US
- Raghava Mutharaju
IIITD – New Dehli, IN
- Axel-Cyrille Ngonga Ngomo
Universität Paderborn, DE
- Stefan Ollinger
Universität Trier, DE
- Alessandro Oltramari
Carnegie Bosch Institute – Pittsburgh, US
- Catia Pesquita
University of Lisbon, PT
- Jay Pujara
USC – Marina del Rey, US
- Michael L. Raymer
Wright State University – Dayton, US
- Ute Schmid
Universität Bamberg, DE
- Luciano Serafini
Bruno Kessler Foundation – Trento, IT
- Cogan Matthew Shimizu
Wright State University – Dayton, US
- Daria Stepanova
Bosch Center for AI – Renningen, DE
- Valentina Tamma
University of Liverpool, GB
- Annette ten Teije
VU Amsterdam, NL
- Riccardo Tommasini
INSA – Lyon, FR
- Frank van Harmelen
VU Amsterdam, NL
- Eugene Vasserman
Kansas State University – Manhattan, US
- Gustav Šír
Czech Technical University in Prague, CZ

Report from Dagstuhl Seminar 25292

New Frontiers in AI for Game Design

M Charity[*1], Michael Cook[*2], and Nicolaas Vas[*3]

1 **University of Richmond, US.** mlc761@nyu.edu
2 **King's College London, GB.** mike.cook@kcl.ac.uk
3 **Billund, DK.** nicolaas_vas@hotmail.com

Abstract

Game design has often influenced, and been influenced by, computer science research. In recent decades researchers and designers have sought to bring these two fields even closer together: to find new ways to think about the game design process; new ways to drive innovation in computer science through playful exploration; and ultimately find new ways to play, design and think about games through computational lenses. AI is impacting the creative industries in more ways than ever before, some welcome, others less so. It is important to find ways for both researchers and practitioners to come together to map out possible futures for this space, to understand where research can contribute, what it can learn from game design in return, and how we can enrich the creative practice of everyone involved.

This report covers Dagstuhl Seminar 25292: *New Frontiers for AI in Game Design*. It outlines the motivations for organising the seminar, summarises many of the working groups that took place, and disseminates some of the games, theories and other materials created during the seminar. The report offers theoretical frameworks, working prototypes and exploratory discussions that present many possible futures for both the creative practice of game design, and the academic field of games research. None of these futures are singularly correct, and many more remain out there to be found; this document merely charts out some possible paths into the unknown that we found exciting to consider.

Seminar July 13–18, 2025 – https://www.dagstuhl.de/25292
2012 ACM Subject Classification Applied computing → Computer games; Computing methodologies → Artificial intelligence; Human-centered computing → Human computer interaction (HCI); Applied computing → Personal computers and PC applications
Keywords and phrases artificial intelligence, Computational Creativity, Game Design, Human-Centred Computing, Procedural Content Generation
Digital Object Identifier 10.4230/DagRep.15.7.124

1 Executive Summary

Michael Cook (King's College London, GB)
M Charity (University of Richmond, US)
Nicolaas Vas (Billund, DK)

The relationship between games research and game design is as old as it is complicated. If we track back through the history of research in and around games, we find many researchers were also designers, and research touching on games was often playful in nature. Similarly, game design is an exploratory creative practice that often asks new questions about technology and art, and in turn drives innovations elsewhere. Today, the overlapping space in between

[*] Editor / Organizer

these two spaces is richer than ever, with many game designers also working as researchers, and vice versa. On a community scale, however, many barriers and gaps remain. Does a researcher have to become a practitioner to impact game design? Does a game designer have to know all the literature and theory to engage with research? How does one start out learning to make games, or understanding how to make sense of existing research?

As a creative practice, game design itself is richer, and more important, than ever. Games are a vast cultural space that has spilled out of consoles and mobile phones into museums, cinemas and schools, and their economic, artistic and social importance has never been clearer. Game design is thus not simply of relevance to a few hundred people working as professionals in large commercial companies, but a skill that is important to tens of thousands of independent creators and artists, and millions of hobbyists all around the world. Roblox, the most popular videogame with young people in America, is a game design toolbox at its core, a many of its millions of players use it for exactly this purpose. As of mid-2025, when this seminar took place, it is estimated that Roblox alone contained nearly 50,000,000 games and interactive experiences – the majority of which have been designed by its younger players.

Games and game design has always been a rapidly-changing and tumultuous space, particularly in the area of digital games. Major technological shifts have transformed the industry almost every decade, from the shift away from arcades, the introduction of personal computers, the transition to 3D graphics, the wider availability of the Internet, the expansion of digital storefronts, and the broadening access to game development tools. One factor that, as we write this, has an unclear future impact on game design is the advent of "generative AI". AI as a general field has a long and rich history with games, and games are one of the key domains that shaped the history of AI into what it is today. The impact of generative AI on design remains unclear, but the split opinions over the technology and its uses also provides a reason for us to re-examine how research and technology in general are used to shape creative fields like game design – what is useful, what is harmful, and how can we include the people most affected in the research process?

All of this brings us to the motivation for this seminar, tentatively titled *New Frontiers in AI for Game Design*. Inspired by some of the past seminars focusing on games research, we wanted to create a week of big ideas, bold experiments and lots of rich discussion and plan-making. Given the practical nature of our chosen topic, however, we also wanted to explore a more practical approach to a Dagstuhl Seminar, centred around play, exploration and application. Our invitee list included a wide variety of backgrounds, including award-winning independent game designers, industry-leading experts on play and creativity, and world-class researchers. Most of our attendees had some experience making games, whether that be physical or digital, alone or in groups, as a commercial product or a freely distributed project. Our hope was that by bringing people with so many distinct – but overlapping – perspectives on design and play, we could explore not only new frontiers for game design, but also new frontiers for how research and practice can talk, share ideas, collaborate and grow closer.

Bringing people together with different backgrounds, experiences and ways of working always presents challenges. It can also be daunting to turn up to a new community and be sure that your contributions will be accepted, or that you will be able to understand your peers' perspectives. We took some steps to help set the tone for the seminar. Nicolaas, one of the organising team, meticulously designed a wonderful icebreaking session where invitees helped each other make their own alternative physical nametags. As part of this, they also got a taste of *Dagstyle*, a creativity system designed by Nicolaas for use during the week, as

Figure 1 Some of the zines we provided at the start of the seminar – including zines about making zines. The purple cameras were used in the icebreaker namebadge-making, but also used throughout the week for projects and recording the seminar. They printed directly onto thermal paper from the camera itself, making them a useful tool for rapid creative projects.

an inspiration and a source of things to hack and repurpose during the week of workshops. Dagstyle is a visual language of sorts, that can be used for many things. An abstract later in this report provides an overview of the system, should any future Dagstuhl Seminar want to use it in their own work. Most of our working groups used Dagstyle in some form or another, with one of the most prominent uses of it in Section 3.2, where it is used as a prototype for a visual programming language.

During the week, we followed the traditions of the previous games research seminars held at Dagstuhl, by inviting working group proposals at the start of each day and allowing people to self-organise into groups. Each group would then report at the end of the day on their work. The reports you will see later in this document record many of the working groups that took place during the seminar, covering topics such as a design language for gameplay curves (Sec. 3.5), an examination of design intent (Sec. 3.6) and a study of generative AI in the context of rapid prototyping (Sec. 3.8).

Generative AI's impact on discourse in both computer science research and on the games industry was impossible to ignore in 2025, and we wanted to ensure people were able to explore and experiment freely during the week, without being trapped in projects they didn't want to be a part of. To fit this, we instituted a "traffic light" system, suggested by Florence Smith Nicholls, which allowed working group organisers to label their group and how they planned to use generative AI for their project, if they chose to use it at all.

A project labeled as "green" meant that generative AI would be the focus or core of the project design. The projects proposed under this category often focused on the use of large language models, text-to-image generation, or other commercial or open-sourced generative AI models as part of the final prototype or in the design process.

A project labeled as "red" meant that the use of generative AI would be avoided entirely for the project design. Many of the projects proposed under this category focused on more analog forms of game design and content creation or the conceptual practice of game design and around user intent and social dynamics of games.

A project labeled as "yellow" meant that the use of generative AI would not necessarily be the focus and could be used agnostically. Many of the projects proposed and completed for working groups during this seminar fell under this category. By defining these clear categories for the directions of the projects proposed, seminar attendees were able to understand which projects they were interested in joining and comfortably set boundaries before committing to a working group.

One of the great joys of bringing together so many creative, generous and kind people is that a community almost instantly forms, and unexpected things emerge from it. In addition to our working groups during the day, our attendees put on a wide variety of social activities in the evening which helped attendees get to know one another better, and to play. Our thanks to Tiago Machado for organising a tango lesson, to Mike Cook and Gillian Smith for leading a livecoding workshop, to Emily Short for running a collaborative game writing night about AI fear, and to Claus Aranha for ending the week by putting together a night of *Slideshow Karaoke*.

We encouraged our participants to think broadly and explore what they were passionate about during the week, particularly as it was an opportunity to work with a unique combination of people and skillsets. However, we welcomed working groups that were centred around prototypes and small working examples, where it made sense for the questions being considered. We were delighted to see many prototypes and interactive projects emerge from the week, including almost a dozen playable games, some of which have already been archived online. Parallel to this, we provided plentiful crafting materials and support for making *zines* in conjunction with working group outputs. Zines are small booklets, traditionally made using cut-and-paste techniques and photocopiers. We encouraged the use of Nathalie Lawhead's *Electric Zine Maker*, and several participants made zines as part of their participation – notably as a key part of the output for the working group on Keepsake Games in Section 3.12. Many of the zines made during the seminar are included as part of this report, including Nicolaas' Dagstyle zine on making zines, which we presented at the beginning of the week.

Figure 2 Our seminar's group photo, captured on one of the cheap cameras we supplied to participants, printed on thermal paper. Thanks to the Dagstuhl staff for taking extra photos for us!

Dagstuhl Seminars often have an otherworldly quality to them – a gathering of people that often seems impossible or improbable, in a beautiful, isolated setting, where all the traditional rules for how we work are thrown out of the window for a few days. Our stated aim was to explore how AI and related technologies might impact game design, and vice versa. Our working groups embraced this challenge in various ways: thinking about new ways to integrate design and theory; finding new applications for established approaches; testing the limits of, and our assumptions about, new technology; and much more besides. The seminar has already given rise to new international collaborations, game projects and funding plans. On a meta-level, though, the seminar *itself* was also an exploration of how research and design can meet in the same place and find play there. It was a week about play that was also playful, where people from different fields, industries and backgrounds could find new ways to communicate, share and collaborate through games and creativity. As organisers, we are incredibly grateful for everyone who came and gave their time, their ideas and their energy to the impromptu community we built over the course of the week.

One of the games created during the week was by Anne Sullivan. Anne brought a wide variety of painting supplies to the seminar, encouraging everyone to use them both for work and play. As part of the working group on *Keepsake Games*, Anne designed a keepsake game specifically for playing while at a Dagstuhl Seminar. The player is randomly given, or chooses, a series of creative prompts and then designs a postcard to give to another attendee at the seminar. The prompts provide suggestions for what to put on the postcard (for example: something representing a working group), and who to give the postcard to (for example: someone new you met at the seminar). We concluded the week by inviting all attendees to create postcards using Anne's game, using the leftover crafting supplies and Anne's brush pens, and exchange them with other attendees before leaving. It was a brilliant opportunity for reflection at the end of the week, and for strengthening the community ties that had been created over the preceding days.

The rest of this report is dedicated to individual reports from working group leaders and others who contributed to the week's activities. There are many attachments in the report in the form of PDFs for items such as zines. Currently, Dagstuhl has no way to officially archive or preserve interactive works, however at the time of writing there are several digital and physical games produced during this seminar which can be accessed online at `http://playdagstuhl.itch.io`. We will endeavour to keep these projects accessible at this URL for as long as possible, and will encourage their authors to upload archival versions to other locations (such as the Internet Archive) as well. We hope you enjoy reading this report, and look forward to returning to Dagstuhl one day to play again.

References

1 The Experimental Gameplay Workshop.
Online: https://www.experimentalgameworkshop.org

Figure 3 One of the postcards made on the final day, in this case by the postcard game's designer, Anne Sullivan. The postcard features patterned designs from *Calico*, a boardgame which was played at the seminar, and a flower that Anne spotted around Dagstuhl itself.

2 Table of Contents

3 Working groups

3.1 New Frontiers in Tamagochi

Claus Aranha (University of Tsukuba, JP), Duygu Cakmak (Creative Assembly – Horsham, GB), Alena Denisova (University of York, GB), Matthew J. Guzdial (University of Alberta – Edmonton, CA), Florence Smith Nicholls (Queen Mary University of London, GB), Yuqian Sun (Royal College of Art – London, GB), and Sabine Wieluch (Universität Ulm, DE)

Digital Pets are a unique form of digital game, blurring the barriers between the physical and digital, the in-game world and the real world. Bandai's "Tamagochi", released over 30 years ago, is probably the most representative among digital pets.

Taking care of make-pretend creatures is a form of play that goes all the way back to the first human children taking care of dolls. The original Tamagochi toy used its hardware and programming to simulate how often it wanted to be fed, to display random events like sickness and boredom, and to give the user feedback about its internal state. As technologies advance, we ask ourselves what new twists they can bring to this very old form of play.

Recently, advances in the hardware and machine learning technologies has led to a revolution in fields such as Artificial Life. Yearly competitions on virtual creatures are held, where computer programs simulate forms of life, including their capacity for reproduction, tending to their base needs, and eventual evolution into new life forms [1].

In this context, we proposed this workshop to discuss how research advances in Artificial Life and Artificial Intelligence could interact with the concept, design and implementation of Digital Pets. And, in the opposite direction, what the experiences of the interactions between humans and digital pets could inform to the research of Artificial Life forms.

In practice, the workshop was divided in two parts: during the morning, the group discussed our experiences with digital pets: What are representative and unique examples; what are their characteristics; what are their design principles. The discussion evolved into forming a loose taxonomy of digital pets, and a conversation about what new directions could be taken in their design. During the afternoon, the group separated into teams to design and prototype new digital pets, based on the discussion in the morning, as well as experiences from past working groups in this seminar.

3.1.1 Discussion on Digital Pets

The first activity of the working group was a brainstorming exercise where the participants came up with representative examples of digital pets. These examples included traditional digital pets such as Desktop Companions, Nintendogs, Digimon Virtual Pet, and Furby; games that included digital pets as extra content, such as Sonic Adventures Chaos Garden; Non-interactive "aquarium" games where the player only observes the digital pet, such as Tabikaeru and Neko Atsume, non-game devices that were inspired by digital pets such as Pwnagochi and Scanner; and non-eletronic "toys" such as Seamonkey kits.

During this brainstorming exercise, the working group listed the kinds of experiences of play that were associated with the idea of digital pets. To start with, digital pets are expected to have needs that must be tended to by the player, such as food and play. By tending to these needs, it is expected that interacting with the digital pet becomes part of

and parallel to the player's life. In this sense, the digital pet has an independent life that interacts with the player's life, and this allows the player a window into a different world, where they can explore themes of change, growth and death.

From this concept of "independent, parallel life", we defined the design goal of a digital pet as the creation of a personal connection with the player. This aligns with characteristics seen in most of the digital pets discussed, such as cuteness and helplessness, which contribute to create this sense of dependency in the player. In this sense, digital pets that include a physical component, such as Tamagochi, can have an enhanced sense of permanence, in that they continue existing even if the digital creature dies. Although it is possible to reset a dead Tamagochi, this sense of connection has led people to perform rituals such as burials on their dead digital pets.

As the discussion came to a close around the idea of attachment between human and digital creature, the working group formed two teams to create prototypes of digital pets. Two projects were proposed: "Dagochi", with the idea to use undirected learning to create life-like digital pets, and "Rocking with Charisma", focused on ideas of physicality and attachment.

3.1.2 Project 1: Dagochi

Dagochi arose from a desire to convert Reinforcement Learning (RL) agents into Tamagochi-like beings. We began by creating a small grid world directly inspired by Little Learning Machines [2]. In this grid world, agents could move over land to try to collect crystals and avoid fire. Collecting all crystals was considered a terminal state, which would grant a large positive reward. Stepping into fire was another terminal state. We implemented a simple tabular Q-learning agent for this environment. While the environment is small enough to perfectly calculate optimal paths, we specifically wanted to engage with the learning process.

To the simple RL environment described above we iteratively added the following features in an attempt to make it more digital pet-like. We added the ability for humans to create their own maps, defined as text files, with "-" representing ground, "C" representing a crystal, "F" representing fire, and "A" representing an agent's starting location. We felt this would allow for more interactivity in terms of humans in-essence metaphorically feeding environments to the RL agents.

We added a population of agents with five initial agents nicknamed: Sally, Brandon, Peach, Dawn, and Don. We called these agents "dagochi". Each rollout would select a single agent, and have that agent attempt to solve a single randomly selected map. We also added print messages to express the RL agent's Q-table as a metaphor for its emotional state. A Q-table with mostly negative values would mean the dagochi was depressed, so even winning the map would leave them saying "Of course it went badly" through a print statement.

We next added the ability for dagochi to breed, with a small chance of this happening if two dagochi in a row ended up on the same randomly selected map. We wrote a simple script to combine the parent names to create a child name, leading to names like "DSally" for the child of Dawn and Sally. Similarly, we added a small chance of one dagochi killing another, which would delete that dagochi's Q-table.

Overall, we found these small additional features surprisingly effective at making the RL agents feel characterful. Figure 4 includes an example output from the system in terms of a screenshot of the terminal running it. In this example we can see one dagochi first have a child with another dagochi before killing them, then exclaiming "Oh wow I did great!". These sorts of easily narrativized moments happened repeatedly while watching the simulation. It felt not unlike watching an aquarium or other collection of pets interacting.

```
Dagochi Dawn: playing heaven!
--- Dagochi Dawn managed to collect 70.0 reward! They stayed alive for 100% ---
--- Congratulations to Dawn and DSally for the birth of their child Dlly! ---
--- Oh no! Dawn just killed DSally! They will be missed. ---
Dagochi Dawn: ٩( ᐛ )۶
Dagochi Dawn: Oh wow I did great!
```

Figure 4 A print out example of one moment in a single run of Dagochi. In this example dagochi Dawn and DSally give birth to a child "Dlly", at which point Dawn kills DSally before exclaiming "Oh wow I did great!"

Based on this experience, we think there's a real potential in making reinforcement learning more approachable and/or developing novel play experience by continuing to investigate the intersection of RL agents and digital pets. Dagochi can be accessed through a github repository[1].

3.1.3 Project 2: Rocking with Charisma

Our main design aim was to create an analogue virtual pet game. We were drawn to the idea of using a found object, as this would allow a player to use a mundane thing and imbue it with life. Inspired by the proliferation of pebbles around Dagstuhl, we settled on the familiar concept of a "pet rock," designing a new solo TTRPG around caring for it.

The game "Rocking with Charisma" has two main phases. The "Becoming" phase involves choosing a rock, attaching (or drawing) an eye for it, and then calculating its two main stats, health and charisma. Any rock starts out with a charisma of 1, but health is determined by rolling a d6 + 2. In the "Care" phase, each carer must perform a daily action to maintain it so that it doesn't lose health. We wanted our game to be a collaborative game, so there can be multiple carers; this means that more carers means there is more opportunity to lose health, however the more people involved, the greater the chance of gaining charisma. This is because at the end of each day, you roll a d20 + n in which n is the number of carers. If you roll 10 or higher the pet rock gains 1 charisma. Each time it gains charisma, it also gains an eye. At max 5 charisma, a rock can have its own pet rock (Fig.5), and the cycle continues.

Figure 5 Left: A pet rock we created as part of our prototyping process. Right: Pet rock of the left rock.

[1] https://github.com/mguzdial3/Dagochi

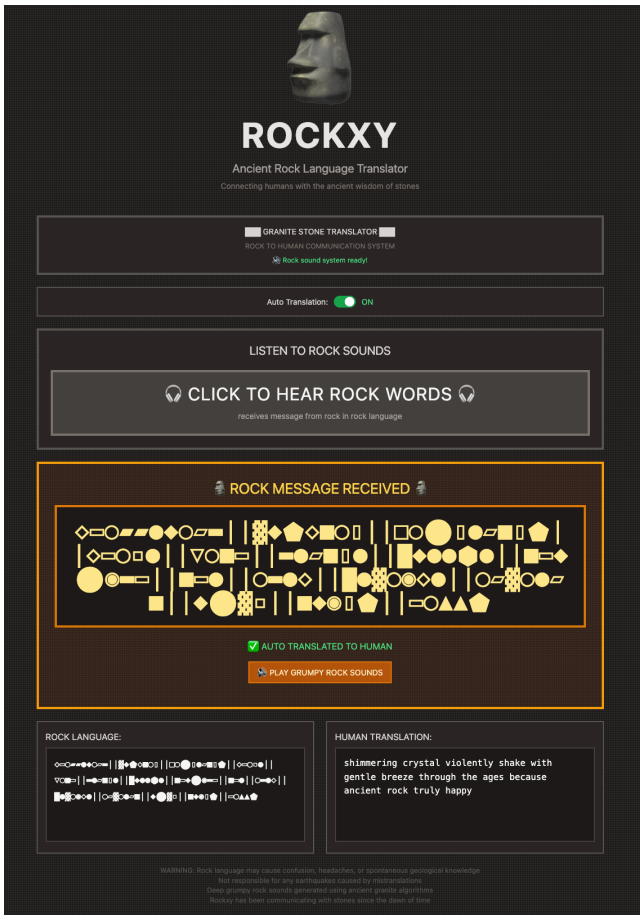


Figure 6 Rock language translator.

Possible daily actions with the rock include, but are not limited to: taking the rock for a walk, decorating your rock, doing a Tarot reading for your rock and designing a passport for your rock. If a rock reaches 0 health it dies, and you should return it to the world. In addition to the analogue game, Yuqian Sun also produced a digital Rock Language Translator [2] that could be accessed by scanning an NFC sticker. This translator produced "grumpy rock sounds" driven by Animalese.js[3], which could then be translated through a fictional granite symbol language, as a way of communicating with your rock. The system maps each letter to unique geometric symbols (vowels as circles, consonants as blocks and shapes), creating a visual cipher that transforms human text into "ancient stone script." Users can either input text to generate rock sounds and symbols, or "listen" to randomly generated rock messages that appear as cryptic symbol sequences.

This prototype is clearly somewhat irreverent in tone, however we believe it provides an interesting provocation on the nature of virtual pets. The original pet rock fad in the 1970s precedes the Tamagochi fad. The selling point was that the rock *did not* require care or maintenance, even to the extent that it is now used as a metaphor for preserving static websites in contemporary digital archiving research [3]. Rocking with Charisma thus imposes the Tamagochi logic onto an earlier, analogue toy.

[2] https://github.com/sunyuqian1997/Rock-Language-Translator
[3] https://acedio.github.io/animalese.js/

3.1.4 Conclusions

The working group discussions, prototype development, and subsequent play and presentation, injected an air of creativity around the ideas of "Tamagochi". Although digital pets are not the first thing that comes to mind when we think of computer games, we think that the play around making connections with make-pretend life form speaks to something fundamental in human nature. This may be at the core not only of digital pets like Tamagochis, but maybe even in our dreams of artificial life forms in fiction and research.

References

1 Sam Kriegman. *Why virtual creatures matter*. Nature Machine Intelligence, Vol 1, page 492, 2019

2 Dante Camarena, Nick Counter, Daniil Markelov, Pietro Gagliano, Don Nguyen, Rhys Becker, Fiano Firby, Zina Rahman, Richard Rosenbaum, Liam A. Clarke and Maria Skibinski. *Little Learning Machines: Real-Time Deep Reinforcement Learning as a Casual Creativity Game*. Proceedings of Experimental AI in Games Workshop, 2023 (EXAG)

3 Martin Holmes and Joey Takeda. *From Tamagotchis to Pet Rocks: On Learning to Love Simplicity through the Endings Principles*. DHQ: Digital Humanities Quarterly 17.1, 2023 (DHQ)

3.2 Visual Representation for Video Game Description Language (V-VGDL)

June Bhartia (Télécom Paris, FR), Michael Cook (King's College London, GB), and Nicolaas Vas (Billund, DK)

The Video Game Description Language (VGDL) [1] is a high-level formalism for rapidly creating complete video games from concise textual descriptions. Originally developed in a previous Dagstuhl Seminar , VGDL enables designers to specify sprites, interactions, and level layouts using symbolic rules, making it a powerful tool for procedural content generation and experimentation. However, its text-based format remains largely inaccessible to non-programmers and those unfamiliar with formal languages.

In this work, we explore a visual extension of VGDL, transforming textual keywords into icons and arranging them spatially to create a more accessible and expressive design medium. By representing VGDL rules, mechanics, and levels through composable images, we aim to lower the barrier to entry for game design while simultaneously introducing new possibilities for creativity, collaboration, and interpretation.

3.2.1 Approach

Our approach began with a direct one-to-one visual translation of an existing VGDL game: Aliens, into a set of icons. Each keyword in the VGDL source was assigned a corresponding image, and these were arranged according to the original code's structure. Figure 7 shows this initial mapping. This prototype allowed us to evaluate readability, expressivity, and compression potential. The translation raised several interesting design questions:

- *1. Compression vs. readability:* How much of the original syntax could be omitted while maintaining intelligibility?

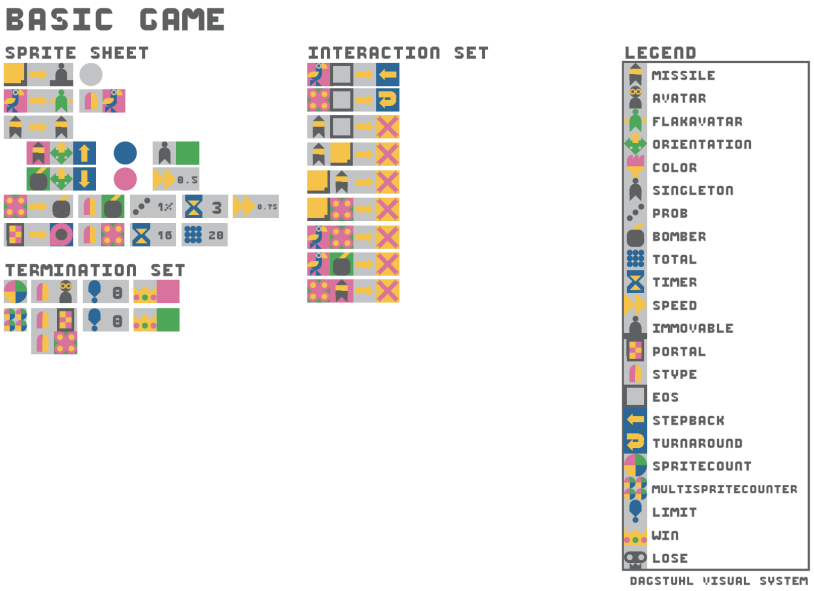


Figure 7 Code for a game, translated from VGDL.

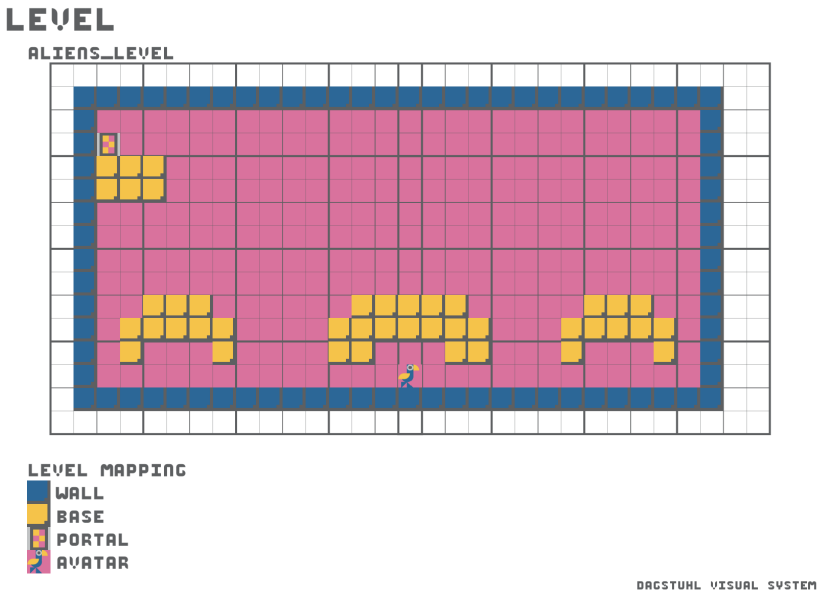


Figure 8 Level for the game shown in Fig. 7, translated from VGDL.

- *2. Layout affordances:* Could spatial arrangements (e.g. crossword-like layouts, scattered clusters) improve memorability or efficiency compared to linear text?
- *3. Expressive gaps:* How might empty space, annotations, or decorations function as part of the design medium?

To answer these questions we attempted to make another prototype using cut up paper icons. We tried to compress it to the size of a postcard and took some liberties with the syntax of VGDL. This prototype is shown in Figure 9.

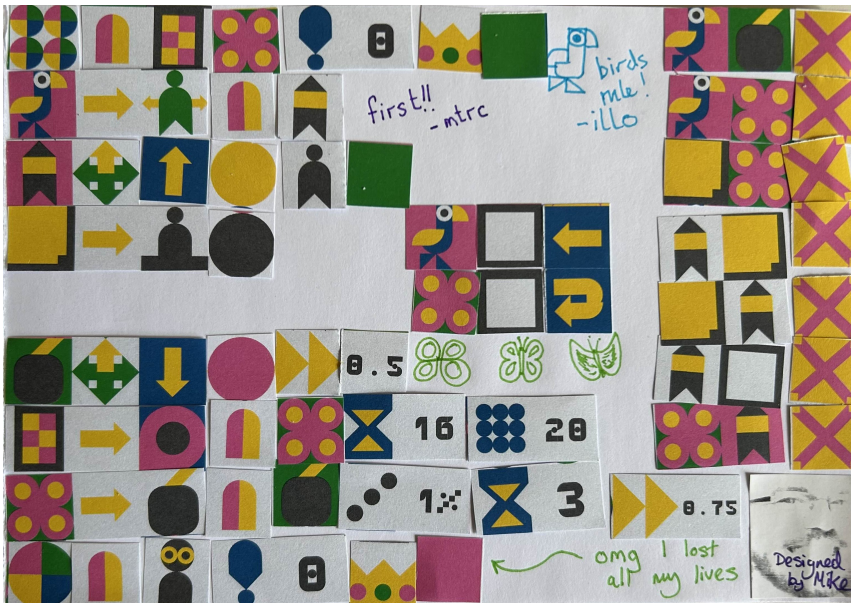


Figure 9 Compression attempt for a game to postcard size, with comments showing expressivity and personalisation.

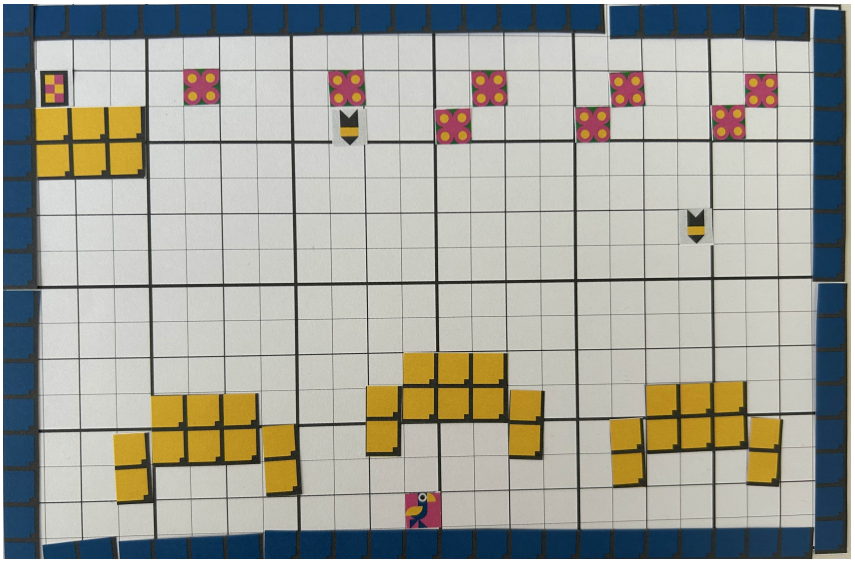


Figure 10 Compressed level for the game shown in Fig. 9.

3.2.2 Discussion and Potentials

Several interesting insights and potentials emerged from our exploration:

1. **Collaborative and Community Play:** A visual system lends itself naturally to physical artifacts that can be shared, such as postcards, magnets, or cards. These could enable community-created games where rules are tangible, remixable, and collectively modified.
2. **Tags, Variables, and Modularity:** Inspired by modular sprite systems, we considered tag-like extensions (e.g. small polymorphic icons) to represent properties and attributes.

3. **Spatial Memory and Arrangement:** The two-dimensional arrangement of icons introduces new cognitive affordances. Crossword-like layouts may enhance recall through spatial memory, though at the cost of efficient space usage. This reflects a trade-off between memory and efficiency.
4. **Cultural Encodings and Localisation:** Visual symbols are not universally interpreted. Cultural context shapes how icons are read, suggesting the need for localisation strategies or multiple representational layers.
5. **Material Affordances:** Unlike text, physical or visual arrangements invite shuffling, cutting, enlarging, or remixing.
6. **Live Coding and Debugging:** Since each visual token has a direct mapping to mechanics visible in gameplay, it becomes possible to highlight active rules in real-time. This creates opportunities for live coding experiences, teaching tools, and interactive debugging.

3.2.3 Future Work

There are many things to still explore around this idea. We have already made a small prototype that can recognize full lines of VGDL from arranged icons. Besides making a digital version of the paper prototype we made during the seminar, there are plenty of directions in which to go, such as live debugging, tangible icons, and exploring automated game design through this.

References

1 Tom Schaul. *A video game description language for model-based or interactive learning.* Proceedings of the 2013 IEEE Conference on Computational Intelligence in Games (CIG), pages 1–8, 2013. doi:10.1109/CIG.2013.6633610

3.3 A Better Mario Kart World

M Charity (University of Richmond, US), In-Chang Baek (Gwangju Institute of Science & Technology, KR), Brian Bucklew (Freehold Games – Walkerton, US), Kate Compton (Vejle, DK), and Matthew J. Guzdial (University of Alberta – Edmonton, CA)

This report details the development of a prototype online multiplayer kart racing game. The focus of this prototype was to create a exploratory and unique kart racer with social multiplayer interaction in an open-world environment via procedural content generation. This work was done by M Charity, Kate Compton, Brian Bucklew, In-Chang Baek, and Matthew Guzdial.

3.3.1 Premise

Mario Kart World (MKW) was released by Nintendo as a launch title for the Nintendo Switch 2 on June 5, 2025[1]. The gameplay – like previous iterations in the Mario Kart series – involves Super Mario characters participating in go-kart races. The tracks include items that can be used against other characters or power-ups such as a speed boost for the kart racer. However, unlike prior iterations, Mario Kart World included a new mechanic of

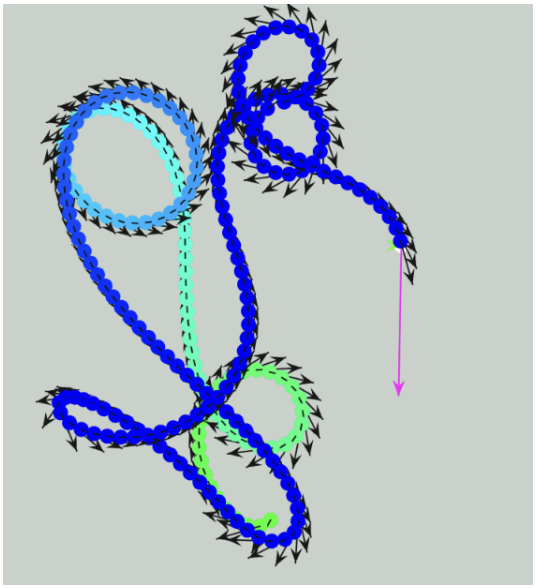


Figure 11 A "karticle" procedurally made track.

open-world driving and navigation. MKW allowed players to freely roam between tracks on the "world map" area and complete minigames or small quests in addition to the traditional track racing.

While the open-world navigation introduced an innovative mechanic for the series – the authors felt the new mechanic did not meet certain expectations of an open world – such as exploration, discovery, and player camaraderie that can be created from spontaneous multiplayer interactions [2]. We developed a prototype multiplayer racing game in 8 hours at Dagstuhl Seminar 25292 in an attempt to explore these missing experiences from MKW.

3.3.2 Development Process

The prototype was originally intended to be a 2.5D billboard style multiplayer game using Mode7 graphics (taking inspiration from the first Mario Kart game on the Super Nintendo system.) The tracks would also be procedurally generated instead of the manually designed tracks used in Mario Kart games. This could allow for more diverse and unique experiences for the player as they drove on the track. See Figure 11 for an example of the procedurally made "Karticle" tracks developed for the prototype.

For the multiplayer capabilities, P5.js[4] and Vue.js[5] were used to handle server and client connections as well as display the graphics via an HTML5 capable browser. We used the p5party multiplayer framework, hosting client relays heroku, and launched the demo publically using ngrok. This demo was made completely free and available to play. When a player connected to the game, they were given a randomly assigned car color and a randomly assigned emoji to represent their character. There were a possible 375 different emojis available for the player – 7 times more characters than the character roster in Mario Kart World. Players had the ability to "honk" at other players connected to the server which would play an audio clip on all players' browsers. This would allow for more direct social connection with other players while they were playing the game.

[4] `https://p5js.org/`
[5] `https://vuejs.org/`

3.3.3 Live Demo

While the Mode7 rendering was unsuccessful for the demo, the top-down view of the track still allowed players to drive and follow the procedurally made track. The server allowed the entire seminar group to connect to the game. With a session size of 30 people this was more than the maximum multiplayer limit of 24 players in Mario Kart World. The prototype did not include power-ups or quests like MKW. In terms of social interactions players could only honk at one another in the game, but playing with everyone in the same room afforded spontaneous interactions external in the game. As such, while we did not accomplish our original objective, we felt that we presented an innovative social player experience in a procedural racing game.

References

1 Mario Kart™ World for Nintendo Switch 2 – Nintendo Official Site. Nintendo. 2025
2 Nathan Gerard Jayy Hughes. Understanding specific gaming experiences: the case of open world games. Diss. University of York, 2023.
3 Brian Shea. Nintendo Says Mario Kart World's 'Value' Justifies Its $80 Price. GameInformer. April 2025
4 J Brodie Shirey. Mario Kart World Devs Explain Lack Of Non-Mario Characters. GameRant. June 23, 2025

3.4 Social Games that You Can Play with Massive Content

Kate Compton (Vejle, DK), June Bhartia (Télécom Paris, FR), Duygu Cakmak (Creative Assembly – Horsham, GB), M Charity (University of Richmond, US), Antonios Liapis (University of Malta – Msida, MT), Tiago Machado (IBM Research – Sao Paulo, BR), Dipika Rajesh (University of California at Santa Cruz, US), and Anne Sullivan (York University – Toronto, CA)

This report details the findings, including a zine and prototype, by June Bhartia, Duygu Cakmak, M Charity, Kate Compton, Antonios Liapis, Tiago Machado, Dipika Rajesh, and Anne Sullivan.

3.4.1 Premise

Like dragons sitting on vast hoards of data, we enjoy an unprecedented wealth of content today.

We have access to cultural content – every historical object from the world's museums, all the posts on Wikipedia, all the assets on Itch.io, or all the slow-burn romances of Archive of Our Own.

We have amassed decades of personal content, a mountain of unsorted photos, social media threads, and algorithmically recorded logbooks of our relationships, playlists, physical locations, GitHub checkins, and heartbeat.

We have generative algorithms, large and small, that create new content on demand. We can create new game maps, text snippets, poems, images, novels, and films, at the press of a button.

Massive data is an important new technology – but humans love to invent new games to play with new technology! So what games can you play with massive content? And why do so many of them involve *other people*? [6]

In this Dagstuhl working group, we looked at many social experiences that use massive content sources, and we've discovered that they have many common design patterns. Below, we will explain a few of these patterns, and examine them in use cases. This paper is far from exhaustive – we believe there is more to be discovered.

3.4.2 Sources, Features, and Uses

To create a playful user experience with massive content, one needs three things:

- **Sources** – you'll need a massive source of content
- **Features** – interaction design patterns for what the users will do with the content, like surfing, annotating, or sharing it with others
- **Uses** – why is interacting with this content meaningful? Are we motivated by the content, or about our journeys and exploration, or is the content just a tool to socialize with each other?

For example, the Library of Babel[7] is a site where every possible book exists, as a reference to the Borges story [3]. Its source is a generator that can create any "possible combination of 1,312,000 characters" on demand. Its primary feature is that each book is represented by a unique URL. Of course, most are random character combinations. There is a small Reddit forum[8] dedicated to finding and sharing occasional discoveries within it, partly as appreciation, and partly for humorous absurdity.

Sources fall into several categories:

- Personal data: Anything we produce in our daily lives, from emails, to photos, personal notes to shopping lists, the music or the crafts we created
- Shared data: where people collectively and intentionally create large amounts of data, for the shared good
- Cultural-collections data: (Museum images, literature) Museums are full of art, digitalized, amazing poems, to literature feasts
- Commercial data: data that is either gathered commercially or scraped from commercial enterprises
- The social data we share explicitly – to be consumed by others in the void – Itch.io
- The social data we share implicitly – the comments to your friends' posts, or some blog that you share for bookkeeping
- Algorithmic generators: data which doesn't *exist* until the moment that it is created

None of these categories presuppose the relationship between the owner or creator of the source and the creator of the game, and the players of the game may have any possible relation to each other. One could make a game for oneself out of one's own data, or be paid by a big-data owner to make a game for others (a common form of digital humanities grant from museums)[4]. Some games are made with data that was not intended (or allowed) to be used for play, which can be a source of both conflict[6] and transgressive power.

[6] For this paper, we use "games" loosely in the "playful experiences" sense used by game philosophers like Bernie De Koven. So for this paper, Pinterest is a game, as is making a mix-tape for your best friend.
[7] `https://libraryofbabel.info/`
[8] `https://www.reddit.com/r/BabelForum/`

Features are the interaction, interface, gameplay, visualization, or sharing features that enable different kinds of exploration and social interactions.

For example, Max Bittker's "River"[1] is a page where players can explore images that people have posted to Are.na. Exploration is done by clicking on one image among many, which causes the page to refresh with other images that are closest to that one in a vector embedding space (more similar). When this happens, the URL changes to the ID of the clicked image. So, if a player finds a favorite "region" of the content, they can share it with others.[9]

We found many such features, and many were common across multiple experiences.

Some features enable or disable ways to navigate through the content. For example, River does not allow **searching**, to instead encourage serendipitous discovery by **proximal exploration**. Exploration can be implemented by many different algorithms and heuristics – similarity-based, curiosity-based, or collaborative filtering. We can direct players to parts of the space near things they like, or direct them to places they have not explored, or to places that *no other user* has explored yet. We can even leave **footprints** to tell them how many other players have been here,[10] or give small UI prizes for discovering new parts of the space.[11]

We can consider each unique page with its "central" image to be a **landmark** that we can share via URL, but it also offers landmarks as **collectibles** and **boundaries** ("Obama eating ice cream," "Heavy Metal Lettering") that players can try to discover. This simple feature (it is just a text suggestion) turns the free-form browsing into a seeking game instead. Landmarks in a large generative space can enable navigation, social conversations, or gameplay challenges.[7][8]

Other features enable players to understand their movement through the space or create **paths and trails** for others to follow. Trails can serve as a memory of how someone has moved through a space or record a path to guide someone else. In River, when M Charity achieved the nearly-impossible feat of navigating to "Obama eating ice cream" they were able to use the browser's back button to demonstrate the path they took, image by image. In experiences like Spotify, we see users curating and sharing unique paths as a new form of content itself, sometimes to take the listener on an unexpected sonic journey and sometimes to make art out of the titles.[12]

We expect **curation sports** to emerge as their own form of play, as users discover particularly interesting spots or become famous as guides or expert-explorers in the space [9]. We also speculated about heuristic **curation pets** as a feature – not monolithic algorithms to create the "best" choices, but "characters" with viewpoints on how to move through the space, like the Tumblr bots mentioned in the case studies below. Could you make a mix tape for a friend? Could you make **a mix-tape-making pet** to give them instead?

Trails, landmarks, heuristics and more can be **reified** into content themselves. Often these are by URL, but can also be embedded in a PNG (as the creatures in the game Spore). This lets the feature engage in the technological ecosystem in any way a URL or image does: A landmark in a generative space could be hidden as a QR code or NFC chip on a sticker or trading card, and physically gifted, swapped, or hidden in a castle!

---

[9] Example regions: `https://river.maxbittker.com/?id=1747330`, `https://river.maxbittker.com/?id=2878636`

[10] `https://www.whatbeatsrock.com/`

[11] `https://neal.fun/infinite-craft/`

[12] `https://www.reddit.com/r/weirdspotifyplaylists/`

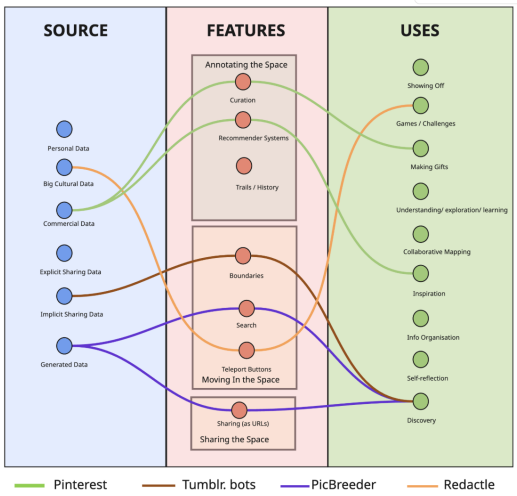


Figure 12 The case studies mapped against sources, features, and uses.

The final component is the **uses** of these experiences. An experience can have any number of features and choose many kinds of content to explore. But what communities and social interactions arise from those choices? What external or artificial structures give them different meanings? We found many such emergent behaviors and social phenomena: showing off, games and challenges, surfing, making gifts, intellectual advancement (understanding/exploration/learning), collaborative mapping, inspiration, creating organized information and annotation (the urge to tidy up a space), self-reflection, and discovery.

When we looked at social experiences, we found two common kinds:

- Collective Experience (sharing the experience) – You go through the experience individually, but share the outcomes with others.
- Collaborative Experience (sharing the space) – You work together in the same environment, creating and contributing to a shared outcome.

3.4.3 Case Studies

Pinterest Boards – June and her friends have developed an annual tradition where they curate Pinterest boards for each other. These carefully crafted collections serve as gifts that keep giving: recipients discover new recommendations and inspirations based on their friends' thoughtful curation choices.

Redactle – This daily word-guessing game presents players with a randomly selected Wikipedia article that has been completely redacted (blacked out). Players must uncover the content by guessing words, gradually revealing the hidden article. The shared daily challenge creates natural conversation points, as friends compare their strategies and discuss the surprising topics they have collectively uncovered.

Tumblr Bots – Automated accounts scan Tumblr posts to identify text strings that correspond to DNA sequences, then post about the organisms they represent. This creates an unexpected treasure hunt where the community celebrates rare biological discoveries hidden within everyday social media content. Users take pride in finding particularly unusual organisms embedded in casual posts.

PicBreeder – This evolutionary art platform demonstrates a complete cycle of generative content interaction:

1. **Generated Data**: AI networks called CPPNs create diverse images
2. **Search**: Users browse and select favorites from random assortments
3. **Discovery**: The algorithm evolves new images by combining and mutating user selections
4. **Sharing**: Each creation has a unique URL, allowing users to share both final images and the complete evolutionary lineage that produced them

When Antonios opens PicBreeder, he encounters a random collection of AI-generated images. By selecting his favorites, he guides an artificial evolution process that produces new images similar to his choices. This allows him to both discover the algorithm's capabilities and share his creative journey with others through the platform's URL-based sharing system.

3.4.4 Further Work

As we discussed these areas, we found example after example that were relevant, each having a different set of features and emerging uses. During the seminar, we also created a zine and an experimental prototype[13] to use the patterns in River for exploring Itch.io games. We hope that this will provide a starting point for others to explore this rich design space.

References

1 Max Bittker. River Notes. `https://maxbittker.com/river-notes`, September 20, 2023.
2 Kate Compton. Liquid Art – A Different Perspective on Generative Art. TEDxNorthwesternU, May 14, 2013.
3 Jorge Luis Borges. "The Library of Babel." In *Collected Fictions*, 1941.
4 Mia Ridge. Playing with Difficult Objects – Game Designs to Improve Museum Collections. Museums and the Web 2011, Science Museum, United Kingdom, 2011.
5 Katherine Compton. Casual Creators: Defining a Genre of Autotelic Creativity Support Systems. Ph.D. Dissertation, University of California, Santa Cruz, 2019. `https://www.proquest.com/dissertations-theses/casual-creators-defining-genre-autotelic/docview/2300563742/se-2`
6 Megan McCluskey. Holocaust Museum Asks Guests to Stop Playing Pokémon Go There. *Time Magazine*, July 12, 2016.
7 S. Risi, J. Lehman, D. B. D'Ambrosio, R. Hall and K. O. Stanley. "Petalz: Search-Based Procedural Content Generation for the Casual Gamer." *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 8, no. 3, pp. 244-255, September 2016.
8 J. Talton, D. Gibson, P. Hanrahan, and V. Koltun. Collaborative mapping of a parametric design space. Technical report, Citeseer, 2008.
9 Iarfhlaith Dempsey. GeoGuessr World Championship 2025 breaks viewership records with waves of community support. `https://escharts.com/news/geoguessr-world-championship-2025-viewership`, September 1, 2025.

[13] `https://github.com/MasterMilkX/zoras-river`

3.5 Dagnamics Description Language

Rémy Devaux (Punkcake Délicieux – Cenon, FR), Claus Aranha (University of Tsukuba, JP), Rafael Bidarra (TU Delft, NL), Emily Halina (University of Alberta – Edmonton, CA), and Gillian Smith (Worcester Polytechnic Institute, US)

Originally setting out to draw out the connection between game mechanic elements and the emotions they inspire, we ended up creating a language which delineates the intended emotional fingerprint of a game. This language can be used to analyse an existing game, or to design a new game.

3.5.1 Introduction

Games make us feel certain ways when we play them. But why? What in a game's mechanics, or in the relations between the mechanics, makes a game feel the way it feels? Can we draw direct connections between mechanics, combinations of mechanics, and emotions, and formulate them in a language? What happens if we then use that language to make a game intended to feel a certain way? These were the questions we had set out to tackle in the workgroup originally titled Expressive Game Mechanics Building Blocks Language.

3.5.2 Moving away from mechanics to get closer to dynamics

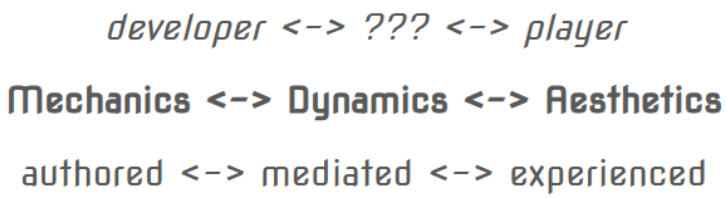


Figure 13 Improvised macro curves.

While discussing our objectives, it became clear that we were particularly interested in the effects of gameplay on the player's emotions, as the gameplay is happening. In the context of the Mechanics, Dynamics, and Aesthetics framework, this means we wanted to concern ourselves with the Dynamics of games. This felt particularly compelling because there are already plenty of tools and methodologies around game mechanics, whereas dynamics are more of a grey area, yet also the gateway to aesthetics, meaning here, broadly speaking, the general impression of the player of a game. Besides, the MDA framework is often criticized for being too mechanically focused, but this may be in part because no-one knows precisely what "dynamics" is, and what happens there. Indeed, the Design, Dynamics, Experience model, which attempts to improve on the MDA framework, also features Dynamics, and leaves them as loosely defined as in the first framework. What definition there is tells us that dynamics are made of the interactions between the game and the player. While playing a game, players experience emotions over time, in reaction to what happens in the game, and in acting on the game through inputs, both, in theory, as designed by the game's designer.

To build off of this, to simplify our approach, and to delineate a helpful methodology, we will assume that players experience games in uninterrupted sessions with no other external stimuli.

3.5.3 Charting dynamics as analysis

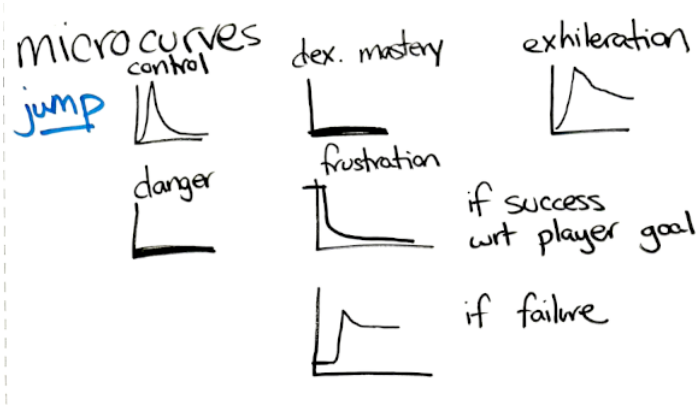


Figure 14 The micro curves for the action of jumping in Sonic The Hedgehog, using 5 different emotions.

Since players' emotions vary over time, it makes sense to draw 2D graphs showing the intensity of an emotion over time as a curve, with different curves for different emotions. These graphs would need to use different scales to provide a complete idea of how the dynamics evolve over the course of a game. We started by deciding on a micro scale, designating the moment-to-moment, like when a player presses a button or is presented with a new information, and a macro scale, representing the duration of the whole game. But we also felt the need for a middleground, and so we decided to also use a meso scale, which designates a short succession of events. For these different scales, we would draw curves for the emotions that seem the most relevant for the game. For example, in Tetris, emotions like anxiety and orderliness might be considered, whereas for a Sonic game, exhilaration and frustration may be more appropriate.

Naturally, playing a game involves more than just two emotions, and so when using the model experimentally to analyse the dynamics of Sonic, we started out with 5 different emotions: control, danger, mastery, frustration, and exhilaration. But when starting by analyzing the micro-scale dynamics of the game, we found that the sense of control followed the same trajectory as exhilaration. And indeed, in Sonic, the two feelings are interlinked, as playing well will usually mean getting a more exhilarating experience. Furthermore, we also found that the senses of danger and mastery did not evolve at all on the moment-to-moment scale. These higher scale notions could in fact be mapped to exhilaration and to the sense of control respectively, only over a wider slice of time. And so we continued our analysis on the meso-scale and macro-scale only considering exhilaration and frustration, and found this to be enough to efficiently pin down the broad dynamics of Sonic.

Additionally, while defining the emotion curves for the meso-scale and the macro-scale, we found it helpful to draw envelopes defining the space in which it would be acceptable for the designers that the player emotions would distance themselves from the emotion curves. For example, a player who is struggling with Sonic's difficulty might get more frustration, but

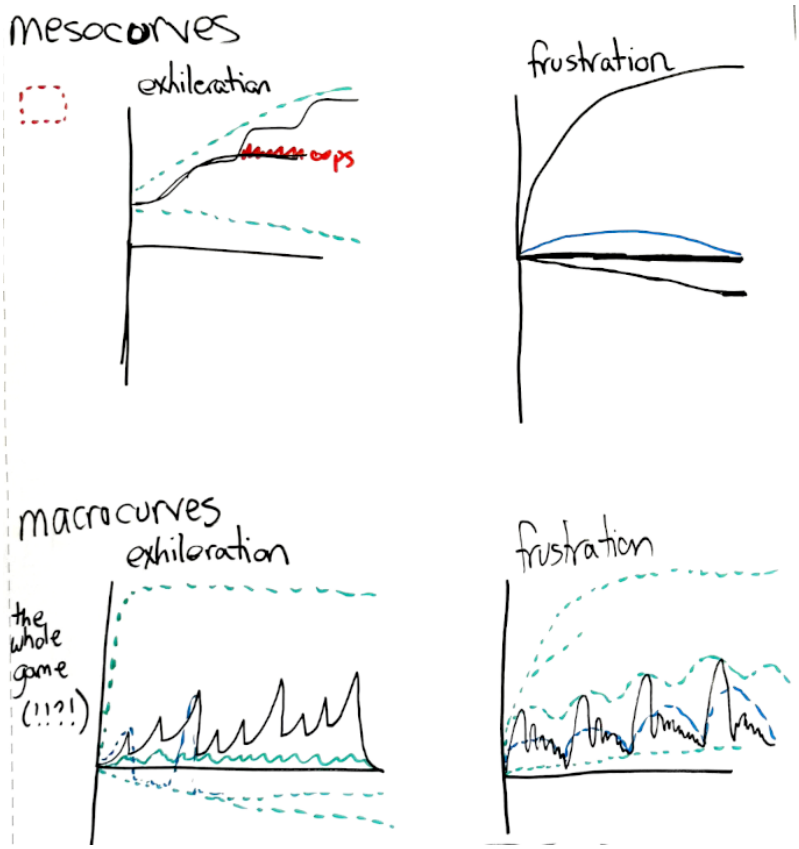


Figure 15 The meso and macro curves for Sonic The Hedgehog, only using the exhilaration and frustration emotions.

ideally only to a point, while their exhilaration may drop down to boredom occasionally to frequently. Conversely, a speedrunner might get little frustration, but would get a sustained high level of exhilaration.

Finally, while drawing all these curves, patterns emerged and we could note a variety of simple shapes coming again and again, and combining in ways that felt expressive. So we established a curve library that could be drawn upon.

3.5.4 Charting dynamics as creative process

Having successfully devised and used a graph-based language to express an existing game's dynamics, we set out to use it in the opposite way, laying out a theoretical game's dynamics, and then making a game that would meet those dynamics. To use the language, we first needed to come up with two emotions around which the game's dynamics would revolve. We felt that those two emotions should contradict in an indirect way to produce interesting dynamics. We came up with playfulness and relaxation.

Trying to decide which curves we should try to draw first, it felt simpler to start with the micro curves for both emotions, and then the envelopes for those same curves. These describe the emotional evolutions that we would want to occur in moments of things happening: for example a player taking action, or a player observing an action taking place, or information being revealed. Even though we didn't yet know what those things would be, we drew two

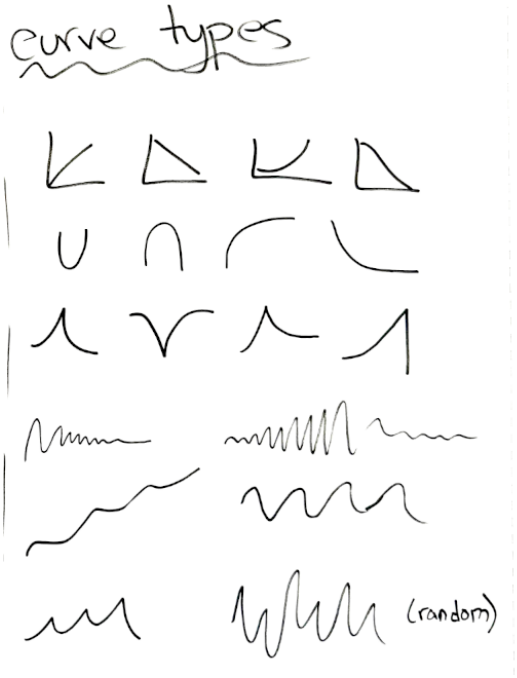


Figure 16 A library of simple shapes and patterns that are very common in emotion curves.

Figure 17 Improvised micro curves.

curves for each emotion. We also didn't shy away from going into the negatives in our graphs, especially for relaxation, on the understanding that this would translate to the opposite of relaxation: excitement.

Figure 18 Improvised macro curves.

After this we drew the envelopes for the macro curves, and then the curves themselves. It felt safer to establish envelopes first, and thus define our desired emotional space, before committing to a more precise emotional experience.

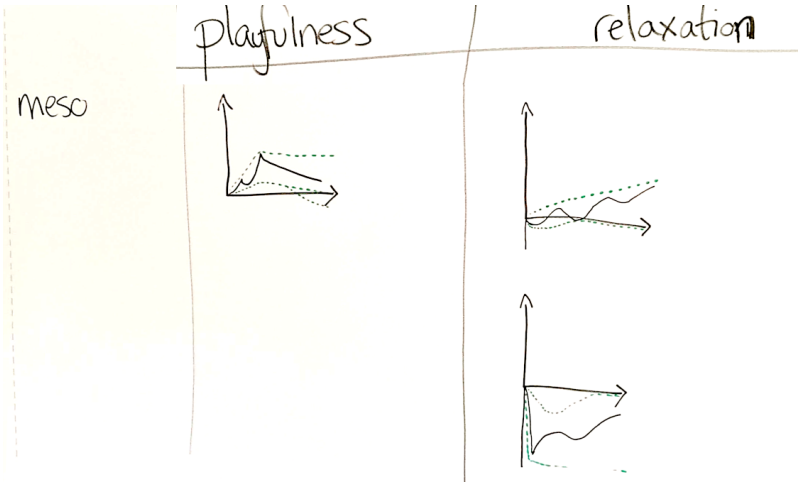


Figure 19 Improvised meso curves.

Then all that remained was the meso scale. Here we went back to drawing curves first and then envelopes. Doing this part last was very interesting because for the whole thing to make sense, the meso curves had to be built off of the micro curves, but also had to describe progressions that would allow us to build the macro curves off of them. Essentially we were making bigger puzzle pieces with our smaller puzzle pieces, with the constraint that the bigger puzzle pieces had to be usable to build the full puzzle.

With our curves all drawn out, it was finally time to design the actual game that would meet them. To make things easier for ourselves, we decided to make an analog game that could be played by two players with what we had on hand. As it happens, we had some coloured pencils and paper, and the emotions we had chosen at the start – playfulness and relaxation – inspired us to make a game with drawing-based mechanics and breathing-based mechanics.

Trying to match the different curves we had drawn, we progressively wrote up a set of game rules, imagining how the game would play out as we went and filling in zones of uncertainty with new rules or changes of rules. Eventually we played the game for ourselves and made additional refinements as we played.

The resulting game is called Breath Checkers and it is a drawing and breathing game where players take turns drawing in gridcells while always taking inspiration from the last thing the other player drew and respecting a checkered pattern in the distribution of grid cells between the players. Playing the game ourselves and observing other people play it, we felt that Breath Checkers managed to match our curves and envelopes fairly accurately, but we were likely biased, since we had just designed the game with that goal in mind. But, perhaps more importantly, we did manage to design a game starting from its dynamics, and the game itself felt novel and fun.

3.5.5 Conclusions and further development

We're very happy to report that this workgroup was successful in creating a new language which bridges game design to player emotions. This language does not establish a direct relation between a game mechanic and the emotions it provokes as originally intended, but it

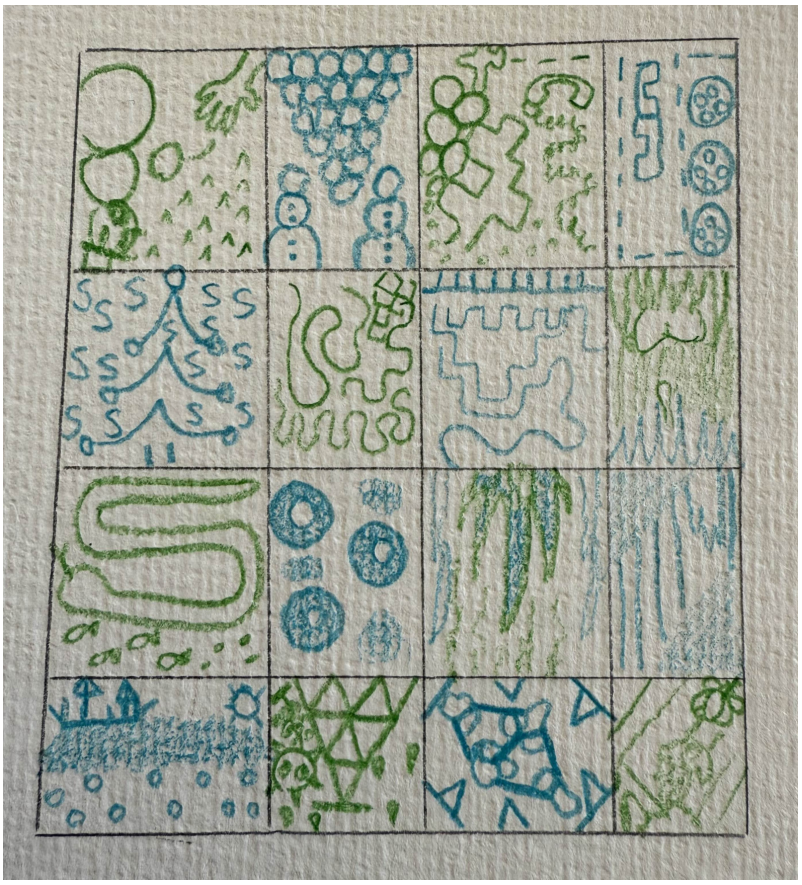


Figure 20 The result of a game of Breath Checkers.

does help analyse and think about how a game impacts its players' emotions, and it also helps laying out strong and workable intentions for creating a game that impacts its players' emotions in a desired way. Additionally, using the language to create a game from scratch was unexpectedly easy. With at least some game design experience, it is fairly easy to intuit what game mechanics might match certain curves. So perhaps drawing a direct link between a mechanic and the emotion it evokes is not quite as interesting anyway.

Designing a game using the language was very inspiring. Interestingly, it felt like we could have come up with a different game with the same emotion curves and this theoretical other game would have had a similar emotional fingerprint. Designing the emotional fingerprint itself was a fun puzzle and surprisingly intuitive. Starting from the micro, meso or macro curve may result in different decisions with the other curves. The starting choice of the emotions one wants to work with is also critical. Finally, this group may not have had much to do with AI, but we believe it did open up possibilities to explore new design space by combining this new language and AI. For example a game prompt generator could use this language, or possibly a new playtesting workflow could be established where players might record their own emotional evolution and an AI tool might do the work or mapping their reported evolution to the correct micro, meso and macro curves. Either way, this language is an exciting new tool for making games.

References

1 Hunicke, R., Leblanc, M. & Zubek, R. (2004). MDA: A Formal Approach to Game Design and Game Research. In *Proceedings of the AAAI Workshop on Challenges in Game AI.* Available from `http://www.cs.northwestern.edu/~hunicke/MDA.pdf`

2 Walk, W., Görlich, D., Barrett, M. (2017) Design, Dynamics, Experience (DDE): An Advancement of the MDA Framework for Game Design. Available from `https://www.researchgate.net/publication/315854140_Design_Dynamics_Experience_DDE_An_Advancement_of_the_MDA_Framework_for_Game_Design`

3 Devaux, R., Smith, G. (2025) Breath Checkers. Available from `https://trasevol-dog.itch.io/breath-checkers`

3.6 Intent: What the heck is it, and how do we measure it?

Emily Halina (University of Alberta – Edmonton, CA), Rafael Bidarra (TU Delft, NL), and Max Kreminski (Midjourney – Santa Clara, US)

Design intent is a topic of increasing interest in the field of games research, with many applications including co-creativity and generative systems. However, our current definitions of design intent are unclear, and vary from author to author. This presents a problem when communicating about or designing systems that incorporate a notion of designer intent, or arguably any co-creative system. In this working group, we set out to better establish potential definitions and ways of measuring intent through observation. After establishing potential definitions of intent through discussion, we ran an informal pilot study observing and interviewing another working group. We present the findings of our pilot study, which indicate the challenges of relying on narrative recollection to determine intent, and give insight into the influence of intent on group working dynamics in a creative context.

3.6.1 Defining Intent

Our working group began with a discussion around potential definitions for design intent. We discussed many definitions, including intent as a realization of a higher level goal, intent as a hierarchical, tree-like structure, and intent as simply "just messing around and finding out." In the end, we settled on a couple of core analogies which helped us to discuss and define intent for the purposes of our pilot study. We consider intent to be an unknown (even to the creator) "guiding force" which drags a designer towards their goal like the pull of a magnetic field. We discussed in depth the "dark matter" of creativity, the off-task activities which actively contribute to the creative process, and how it is closely related to the unknowable true intent we wanted to get closer to measuring.

3.6.2 Measuring Intent

In order to get closer to measuring a notion of true intent, we came up with two potential avenues of measurement. The first is the discernment of intention through interaction directly. This would entail the design of a system intended to elicit intent, then change the underlying system based on a modeling of this intent gradually over time. The second was the discernment of intention through observation. In particular, this entailed a combination of observation and semi-structured, post-hoc interviews with either individuals or members

of a group working on a creative project. The hope is that through a combination of both observation, think-alouds, and post-hoc clarification, we could somehow piece together an accurate portrait of a group member's intention behind their creative decisions. While the first approach seemed promising to the group members, we decided that due to time constraints we would settle on the second for the remainder of the day.

3.6.3 Pilot Study Setup

For our pilot study, we decided to embed ourselves (with permission) into another group for the afternoon. In particular, we chose to collaborate with the *New Frontiers in Tamagochi* group, as they were just beginning multiple creative projects at the start of the afternoon when we joined. We split into two subgroups: one observing the Dagochi group creating a multi-agent reinforcement learning environment, and the other observing the creation of the Rocking with Charisma TTRPG. For more details on the contents of these projects, please see the *New Frontiers in Tamagochi* section.

We observed each of these groups for roughly 90 minutes through their creative processes, then proceeded to conduct semi-structured interviews with each group member to ask follow-up questions about the reasoning behind certain decisions. After this interview process, our group re-convened to discuss our findings, and determine if the groups shared any characteristics.

3.6.4 Findings and Takeaways

The major finding of our analysis was that it is very difficult to intuit intention from just observation alone. In fact, it could be argued that we learned more about creative group dynamics than about the ability to discern any notion of intention behind each group member's actions. In particular, we identified three major takeaways from our observations of the two groups.

The first is that aspects of each member's initial intent were hidden to other group members. While group members each came into the project hoping to achieve something specific from the afternoon of creative work, those intentions were not necessarily communicated or shared among the entire group. For example, one group member who was particularly technically focused had a technically focused retelling, which matched their initial intent. The second is that ideas from group members tend to recombine in different, unexpected ways. For example, different mechanics in the Dagochi project such as breeding and killing emerged from group members' preconceived notion of which mechanics were "obvious" given the multi-agent nature of the environment. The third is that chronology becomes fuzzy and unreliable throughout narrative recollection of events. This was very apparent in the Dagochi project, where each group member gave a different point in time regarding the inception of certain mechanics.

We believe there is still a lot of work to be done toward understanding, interpreting, and measuring intent. While this working group represents a very small step towards that understanding, we believe this problem is increasing relevant towards the design and analysis of co-creative systems and the co-creative process.

3.7 The World Needs Expressive Range Analysis!

Max Kreminski (Midjourney – Santa Clara, US), In-Chang Baek (Gwangju Institute of Science & Technology, KR), Rafael Bidarra (TU Delft, NL), Alexander Dockhorn (University of Southern Denmark – Odense, DK), Emily Short (Oxford, GB), Gillian Smith (Worcester Polytechnic Institute, US), Nicolaas Vas (Billund, DK), and Sabine Wieluch (Universität Ulm, DE)

Generative AI is often presented as something wholly new: a radical break from tradition that introduces opportunities and difficulties unlike any reckoned with before. However, to researchers who study *procedural content generation* (PCG), many of the difficulties surrounding present-day generative AI appear rather familiar.

These common *problems of generativity* faced both by users of traditional PCG pipelines and large pretrained generative models are many and varied. Users may struggle to wrap their heads around a generative system's tendency to produce noticeably homogenous outputs [1, 5], or its inability to produce certain kinds of outputs at all [3]. They may immediately accept the system's first passable output as "good enough" instead of continuing to strive for a better outcome [20], or feel artificially constrained by system capabilities [12]. The impacts of specific parameter adjustments on system outputs may be difficult for users to anticipate [15, 19]. And the difficulty of drawing reliable conclusions about a generative system as a whole from a small number of concrete outputs may lead users to treat these systems as magic [7], rather than as biased cultural artifacts shaped by data curation processes or algorithmic materials with a particular characteristic grain.

We argue that many of the problems of generativity stem from the fact that generative systems instantiate a (frequently vast) *expressive range* of possible outputs, but an individual user will generally only be able to formulate their view of a system in light of a relatively small number of individual artifacts sampled from this range. To mitigate these difficulties, PCG researchers have long made use of *expressive range analysis* (ERA) [15]: a family of techniques used to visualize, reflect on, and make sense of a generative system's entire expressive range. However, several of the assumptions made by traditional PCG research about generative pipelines do not apply neatly to the large pretrained generative models of the current genAI boom [2]. As a result, ERA has not yet seen much application to the problems of generativity in their newly and greatly expanded form.

We convened this working group to identify new potential application areas for ERA; survey recently proposed expansions to ERA as a method; characterize the pain points limiting adoption of ERA in new contexts; and set out a preliminary agenda for translational research intended to broaden ERA's applicability.

3.7.1 What is ERA?

Expressive range analysis in its usual form involves the generation of a very large number of output artifacts from a single generative pipeline; the characterization of each generated artifact in terms of a set of domain-specific, computationally assessable metrics; and the visualization of the distribution of generated artifacts as a set of two-dimensional heatmaps, with each heatmap showing the distribution of generated artifacts in terms of a particular metric pair. The example output of a typical ERA process can be seen in Figure 21. Several

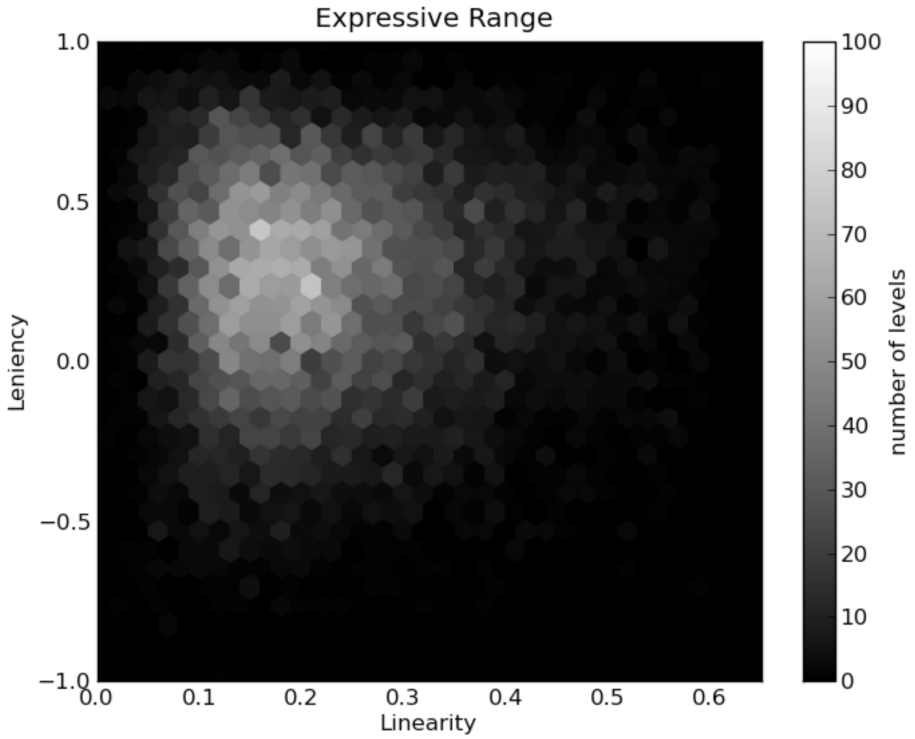


Figure 21 A visualization of the expressive range of a game level generator in terms of two game level-specific metrics, *linearity* and *leniency*. Reproduced with permission from [15].

extensions of ERA have also been proposed, for instance to support analysis of interaction dynamics in PCG-based mixed-initiative co-creation [11] and to improve the utility of the generator fingerprints captured by ERA in various ways [17].

Despite ERA's broad uptake within PCG research, the question of how to define or select appropriate metrics for a particular class of artifacts remains an open problem [18]. Additionally, although some attempts have been made to integrate ERA directly into graphical game creation tools like Unity [6], few existing interfaces allow users to employ ERA without reimplementing all of the necessary components themselves.

3.7.2 Who needs ERA?

We envision several potential user archetypes that might benefit from application of ERA:

- The **computational author**, who aims to craft a generative pipeline that achieves an envisioned output distribution or aesthetic outcome.
- The **analytical researcher**, who aims to better understand the expressive range of *other* people's generators (e.g., to assess biases in large pretrained generative AI models).
- The **educator**, who seeks to broaden public understanding of generative pipelines as opinionated and probabilistic rather than unbiased and oracular.
- The **exploratory tool user**, who plays with generative pipelines in an exploratory or process-focused way but is not directly motivated by the desire to achieve a comprehensive understanding of a particular possibility space.
- The **outcome-attached tool user**, who attempts to use generative pipelines to achieve specific types of desired output but is only instrumentally interested in what the pipeline as a whole is capable of producing.

The boundaries between our envisioned user archetypes are not firmly fixed: people may gradually slide from one category to the next as their interests and needs evolve. For instance, an outcome-attached tool user may become more interested in making tool outputs that stand out from what they've generated before; an analytical researcher may try to address weaknesses they've previously identified in existing generative pipelines by developing new ones; and a computational author may decide to assert more direct curatorial control by selecting and publishing only a few specific artifacts from their generative pipeline's full expressive range.

3.7.3 What do they need from it?

One of the main bottlenecks preventing broader application of ERA is the limited availability of appropriate, computationally assessable metrics that can be used to characterize artifacts in specific creative domains. This has manifested in the past in several important ways:

- Over-indexing in the published literature on narrow sets of metrics defined in prior work (e.g., linearity and leniency for game level generation)
- Non-extension of ERA to new artifact domains, due to the difficulty of coming up with an initial set of domain-specific metrics for a totally novel domain
- Overreliance on metrics that are easy to computationally define, rather than those that capture key aesthetic properties of relevant domains
- Non-application of ERA by potential users other than computer science researchers, due to the difficulty of encoding intuitively salient metrics as procedural code

Notably, many of our envisioned user archetypes will not necessarily have a strong sense of what metrics they're interested in before interacting with a generative pipeline; some of them may not know how to code; and even those who both know what metrics they're initially interested in *and* know how to code may struggle to meaningfully formalize the metrics they care about. How can we support these users?

We believe the answer may take the form of an approachable workbench for conducting expressive range analyses, with built-in support for the definition and iterative revision of "sketchy" example-based metrics, as well as the publication, retrieval, and adaptation of metrics defined by the community of workbench users. Such a workbench might be modeled on Wekinator [8], an approachable toolkit for the example-based definition of simple domain-specific classifiers and other machine learning models for artistic use cases (e.g., novel musical instrument design). Metrics might initially be defined in terms of "general-purpose" embedding or language models (as in, e.g., Luminate [16] or Patchview [4]); refined through the specification of additional examples; and perhaps translated into a more domain-specific ML model or explicit procedural function as the user develops a more specific sense of what they want their metric to capture.

3.7.4 Case studies

Our working group attempted to apply ERA in several new contexts: to the outputs of a pre-existing textual expansion grammar, with new candidate metrics defined in terms of LLM queries rather than procedural code; to a character generator building on the Dagstyle visual language introduced earlier in the seminar; to a large number of *Dungeons & Dragons* oneshot scenario concepts generated by the large language model Claude Sonnet 4; and to an automatically generated space of possible playthroughs of the storylet-based interactive narrative *Bee* [14]. In each case, we encountered new difficulties that might need to be addressed in the extension of ERA.

3.7.4.1 LLM-based metrics for traditional procedural text

The easiest ERA metrics to implement for any given class of artifacts tend to be *syntactic*: focused on surface-level details of the artifact's structure, such as a poem's length in words or the percentage of tiles in a game level that are walls. However, the artifact properties that are most interesting to a human observer tend to be *semantic*: focused on aspects of an artifact's deeper meaning, such as a poem's engagement with a particular theme or a game level's difficulty. To aid the implementation of semantic metrics, we wanted to assess whether LLMs or other open-domain interpretive models might perform well at characterizing flexible aspects of artifact semantics in new domains. As a pilot of this approach, we defined LLM prompts to assess two different dimensions of interest in the outputs of a textual generative grammar for fictional travel guide entries (*evocativeness* and *absurdity*) and employed these prompts as metrics to conduct an ERA of the grammar under inspection.

We found that the LLM did in fact produce seemingly reasonable values for the example outputs we sampled. However, the relatively high computational cost of a single LLM call makes LLM-based metrics harder to apply at very large scales than many traditional ERA metrics, and LLM nondeterminism means that the same prompt may yield a different assessment of the same input artifact on subsequent runs, adding uncertainty to ERA outcomes. These difficulties may become priorities for future work.

3.7.4.2 Image content metrics for Dagstyle characters

Beyond using LLMs to assess the semantics of text, we also wanted to experiment with using open-domain models to assess the semantics of visual content – e.g., representational character icons in the seminar's Dagstyle visual language (Figure 22). Because important aspects of visual experience are sometimes hard to express linguistically, and because vision language models tend to be even more computationally expensive than their LLM counterparts, we decided to pursue an example-based approach to specifying metrics via embedding similarity in a joint text/image embedding space (e.g., SigLIP [21] embeddings). Under this approach, we decided to initially define the semantic characteristics we believed we were interested in as text alone; embed the resulting text strings, and the generated characters we were interested in characterizing; use embedding similarity measures to initially visualize the expressive range of the character generator in terms of our candidate metrics; and then refine the specification of our semantic metrics by redefining the anchor embeddings that defined each metric in terms of specific exemplary generator outputs.

In practice, we couldn't get SigLIP working for crossmodal comparisons in the limited amount of time we had during the working group. This initial failure accentuates the importance of putting together a streamlined workflow for these kinds of comparisons, so that setting up a Python environment capable of running moderately complicated ML models isn't a hard requirement blocking deployments of semantic ERA. Ultimately, we were still able to conduct an ERA of our character generator via purely syntactic metrics (e.g., involving pixel color ratios), but – as expected – these syntactic metrics did not help us make much sense of whether our generator was succeeding or failing at generating a wide range of perceptually different representational icons.

3.7.4.3 LLM-generated D&D oneshots

Next we turned to the use of ERA to characterize the outputs of genAI models – e.g., *Dungeons & Dragons* oneshot concepts generated by Claude Sonnet 4. Since many users are already making use of LLMs for ideation in a tabletop roleplaying context [1], and since

Figure 22 Example characters produced by our character icon generator.

oneshot concepts are individually relatively small, we felt that this might be a good test domain to probe LLM biases and gauge how homogenous an LLM's attempts at creativity might be in a realistic usage scenario.

However, we rapidly ran into a major problem with this approach: when using a single fixed prompt to generate oneshot concepts one at a time, the specifics of the resulting concepts turned out to be so similar that almost no meaningful semantic variation between them was apparent. Prompting the model to generate an entire batch of concepts all at once yielded better within-batch variation, but across multiple batches, the same ideas – down to specific character and location names – continued to show up again and again.

The most promising solution that we could identify to this problem involved the deliberate permutation of the *input prompt* used for scenario generation: rather than generating each batch of oneshots with the same prompt, we could first instruct the model to generate a wide range of different ways to *ask* for a batch of oneshot ideas (e.g., with different phrasing or different implied user personas), then generate a batch of responses using each of these different prompts. This yielded enough semantic variation to merit study, but presents a further conceptual problem by blurring the boundaries of the generative system under evaluation. When using ERA to evaluate an LLM, the input prompt provided to the LLM can be viewed as part of the generative pipeline; treated as an extraneous parameter; or varied lightly to represent a *class* of prompts that might reasonably be employed in a particular usage context – but none of these approaches are obviously, categorically correct in our eyes.

3.7.4.4 Interactive narrative playthroughs

Although interactive narratives have sometimes been framed as generative systems [10] instantiating a "story volume" of possible valid storylines [9], it has so far proven difficult to effectively apply ERA to the analysis of spaces of possible interactive narrative playthroughs. Playthroughs tend to be individually complex, making each one hard to usefully summarize as a small number of easy-to-calculate syntactic metric values. Moreover, the temporal structure that makes playthroughs easy to visualize as individual storylines also complicates the simultaneous visualization of many playthroughs at once, as the storylines tend to become visibly "tangled".

Our workgroup explored several possible approaches to playthrough visualization. The most immediately promising approach involved the simulation of many possible playthroughs of the storylet-based interactive narrative *Bee*. These playthroughs could then be annotated according to the values of certain especially important story state variables at key moments (e.g., the playthrough's end) and visualized on a two-dimensional plane via UMAP dimensionality reduction [13], with the visualization highlighting clusters of similar potential player experiences. Questions we encountered in the process included whether to treat entire playthroughs or individual *moments* from these playthroughs as the artifacts under ERA; whether to use explicitly tracked story state variables, syntactic playthrough properties (e.g., how many times a given storylet has been revisited) or semantic playthrough properties (e.g., the protagonist's inferred level of emotional well-being) as metrics; and how to handle the time dimension of playthroughs in visualization.

References

1 Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 413–425, 2024.

2 Josiah Boucher, Gillian Smith, and Yunus Doğan Telliel. Axes of characterizing generative systems: A taxonomy of approaches to expressive range analysis. In *Proceedings of the 11th Experimental Artificial Intelligence in Games (EXAG 2024) Workshop*, 2024.

3 Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. Nine potential pitfalls when designing human-AI co-creative systems. *arXiv preprint arXiv:2104.00358*, 2021.

4 John Joon Young Chung and Max Kreminski. Patchview: LLM-powered worldbuilding with generative dust and magnet visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024.

5 Kate Compton. So you want to build a generator. `https://galaxykate0.tumblr.com/post/139774965871/so-you-want-to-build-a-generator`, 2016. Accessed: 2025-09-24.

6 Michael Cook, Jeremy Gow, Gillian Smith, and Simon Colton. Danesh: Interactive tools for understanding procedural content generators. *IEEE Transactions on Games*, 14(3):329–338, 2021.

7 Shipi Dhanorkar, Christine T Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. Who needs to know what, when?: Broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 1591–1602, 2021.

8 Rebecca Fiebrink, Dan Trueman, and Perry R Cook. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2009.

9 Jason Grinblat. Emergent narratives and story volumes. In *Procedural Generation in Game Design*, pages 199–207. AK Peters/CRC Press, 2017.

10 Matthew Guzdial, Devi Acharya, Max Kreminski, Michael Cook, Mirjam Eladhari, Antonios Liapis, and Anne Sullivan. Tabletop roleplaying games as procedural content generators. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, 2020.

11 Max Kreminski, Isaac Karth, Michael Mateas, and Noah Wardrip-Fruin. Evaluating mixed-initiative creative interfaces via expressive range coverage analysis. In *IUI Workshops*, pages 34–45, 2022.

12 Jingyi Li, Eric Rawn, Jacob Ritchie, Jasper Tran O'Leary, and Sean Follmer. Beyond the artifact: power as a lens for creativity support tools. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

13 Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

14 Emily Short. Bee. `https://inthewalls.itch.io/bee`.

15 Gillian Smith and Jim Whitehead. Analyzing the expressive range of a level generator. In *Proceedings of the 2010 Workshop on Procedural Content Generation in Games*, 2010.

16 Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. Luminate: Structured generation and exploration of design space with large language models for human-AI co-creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024.

17 Adam Summerville. Expanding expressive range: Evaluation methodologies for procedural content generation. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, pages 116–122, 2018.

18 Oliver Withington and Laurissa Tokarchuk. The right variety: Improving expressive range analysis with metric selection methods. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, 2023.

19 JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

20 Irene Zanardi, Shana Dedò, and Monica Landoni. On the good-enough effect: Children reflect on their AI-generated portraits. In *Proceedings of the 24th Interaction Design and Children*, pages 154–167. 2025.

21 Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

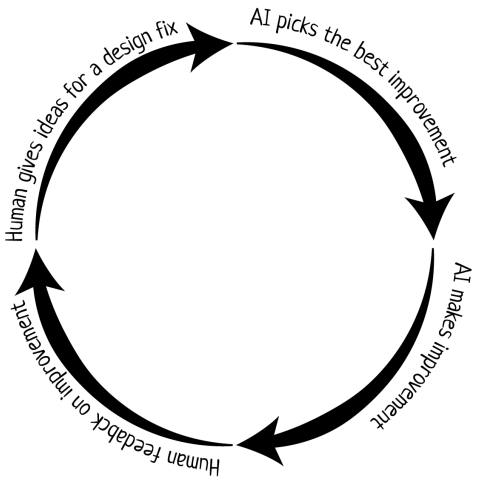


Figure 23 Process followed to produce the games: the human designer primarily tested the game and provided feedback, while the AI selected what (and how) to change in the game code.

3.8 A Game in a Day

Antonios Liapis (University of Malta – Msida, MT), Maren Awiszus (Viscom AG – Hannover, DE), Alexander Dockhorn (University of Southern Denmark – Odense, DK), and Timothy Merino (NYU – New York, US)

While the generation of game content and even complete games [3, 11] has been well-established in academia [15] and practice [16], the now-ubiquitous techno-optimism of data-driven Generative AI or GenAI[14] raises timely and critical questions. It is easy to dismiss claims made on creators' blogs [1] and social media about GenAI's ease-of-use in creating game code or assets; similar claims on nearly every type of work (creative or not) abound. It is arguably easy to dismiss academic research in (partial) game generation via GenAI [14, 17] as proof-of-concept without general applications. And yet, AI-generated content is increasingly popular in published games: almost 20% of all games released on Steam in 2025 have disclosed use of "AI Generated Content", eight times as many as in 2024 [8]. The implications of GenAI in game development practice can not be ignored.

This working group identified several questions around the technical feasibility of GenAI-based games, and their implications in terms of ethics, ownership, and (human) creativity. The main questions investigated were:

- How does GenAI handle the creative decision points of game development, and what does this mean for the creativity of a human developer working with it?
- How closely can GenAI replicate an existing game without access to its codebase, and what are the implications for Intellectual Property protection?
- How "in control" does a human designer with a clear idea for the game feel when implementing it via GenAI?

[14] We primarily consider GenAI to cover Large Language Models (LLMs) and Text-To-Image Models trained on massive amounts of data, often as black-box models owned by corporations.

Primarily, the working group took the challenge of creating fully functional games in a day or less, using as little human input as possible. The practical activities revolved around an iterative loop of the human giving ideas for fixes and improvements to the work-in-progress game, the AI picking the best improvement and implementing it in the game, then the human giving feedback on the improvement (see Fig. 23). The process followed is a special case of vibe-coding [5], which is the dominant approach for using LLMs for coding tasks. The practical game development tasks described above were complemented with ad-hoc points of reflection, discussing with the group how the human "creator" was feeling – especially in terms of ownership, creativity, and control over the process and the product [6].

3.8.1 Case 1: GenAI handling [most] creative decisions

For Case 1, we wanted to leave the maximum creative freedom to the AI, focusing only on our subjective comments about the game's playability rather than about the creative decisions. To maximize the end-to-end ideation, we followed a game jam format [7] and generated one theme using an online (non-LLM) theme generator: the resulting theme was "islands". After this, we relied on Anthropic's Claude as our LLM of choice for all concept and code generation phases.

Using the generated theme "islands", we prompted Claude for game ideas for this theme "that will wow the other participants", which returned 6 game ideas. We then asked Claude to rank these ideas on which "will generate the most buzz"; the response ranked *Island Genesis* first, which we then used for the rest of the game generation pipeline. The original description of *Island Genesis* in Claude's first response for Case 1 was:

> *A reverse city-builder where you play as a volcanic island that's slowly sinking. You must strategically grow land masses and guide the last inhabitants to safety before you're completely submerged. Time pressure with emotional storytelling.*

Following this description, we followed the iterative cycle of Fig. 23 to produce a game using first the PICO-8 [9] engine and then the Godot [12] game engine.

Using the PICO-8 engine, the first LLM response already added the very basics of the game (see Fig. 24; top left): a visualization of an island, the people to be saved, and a basic game state display; the island also already slowly sinks. However, cursor interaction did not work and the game was unplayable. Follow-up iterations involve a human playing the game and producing feedback, with Claude listing possible improvements and then picking the (LLM-perceived) best one to work on. Improvements over ten iterations, lasting around 2 hours, resulted in visual polish (e.g. particle effects, progress bars) and game design additions (e.g. better villager pathfinding, slowly renewing materials used for growing land). The resulting game is playable (see Fig. 24; top right) and, at times, even engaging. The visuals are polished enough for a PICO-8 game.

Following the relative success of the PICO-8 *Island Genesis*, we tried replicating the process (with the same high-level description) with the Godot [12] game engine, which is more complex. Given the same time and effort, the resulting game (see Fig. 24; bottom) is much worse than the PICO-8 game. Likely reasons include the more complex visuals (high-resolution textures) and complex file structures (compared to the single script used in PICO-8). The former increases the expectation for visual quality and increases the challenge of image generation, while the latter poses challenges for the LLM to process (given limited context length) and to create responses for (which takes far more time than for PICO-8). This suggests that a game with a much larger scope – regardless if that scope comes from more complex software or more complex game mechanics – will result in worse results using only GenAI planning and programming.

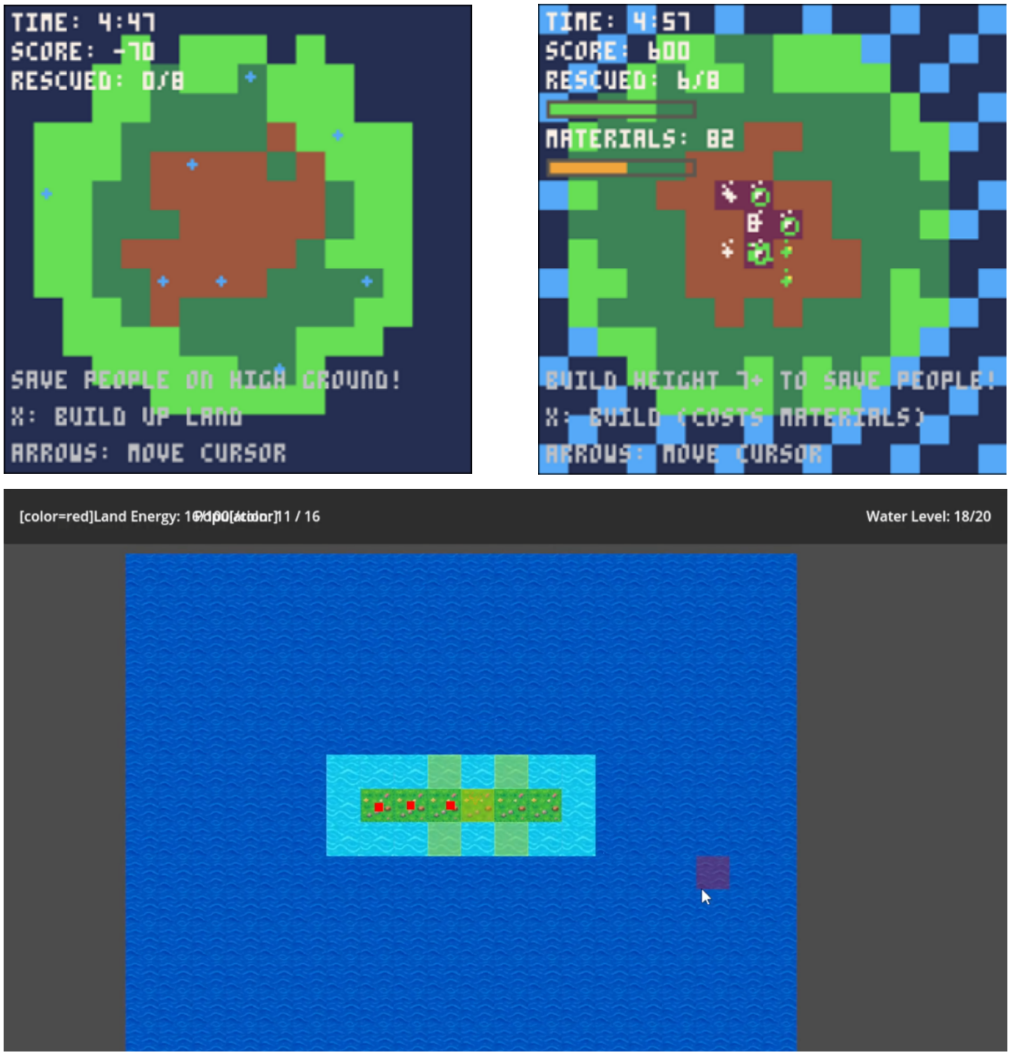


Figure 24 Screenshots of *Island Genesis* in PICO-8 (top) and Godot (bottom).

3.8.2 Case 2: GenAI reproducing an existing game

While the other two cases explored the balance of creative decisions between LLM and human co-creators, Case 2 explores how this game development process works when neither human nor LLM have creative agency. All creative decisions are already taken in advance, by someone else: in our case, the creator of an existing game. Another benefit of this approach is that we already have available the description, screenshots, and access to the original game to play and assess intended gameplay when giving feedback to the LLM. This is unlike the other two cases where, to use the creative journey as a metaphor, neither LLM nor human know the destination (instead, they explore together) and it is unclear what constitutes the end of the journey.

Given the success of Claude at generating PICO-8 games in Case 1 (see Section 3.8.1), we browsed recent high-ranking PICO-8 games on *itch.io* and chose the game *SuperHotRoids*[15] to replicate. Among our selection criteria, we considered that the game's recent release would make it less likely that the LLM is trained on information about the actual game.

[15] `https://ghettobastler.itch.io/superhotroids`

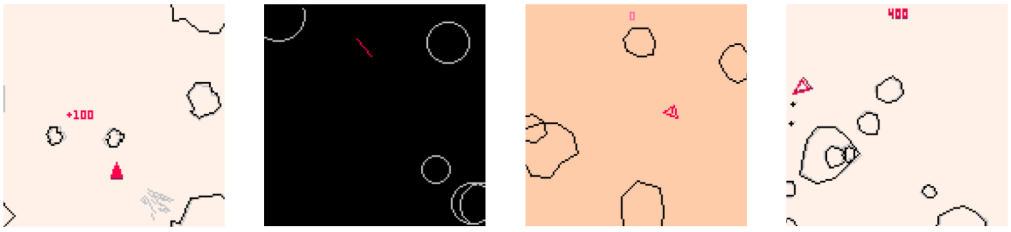


Figure 25 The original *SuperHotRoids* (far left) and checkpoints of the reproduction process at 1, 5 and 30 queries.

SuperHotRoids mixes up the mechanics of *Superhot*[16] and *Asteroids*[17]: a spaceship attempts to shoot down asteroids while time slows when the player does not interact with the game's controls. We will use the title *SuperHotRoids* to refer to our own GenAI-made game in Case 2, since we try to replicate it in full.

Given an initial description of the game mechanics and screenshots from the original game, we prompted OpenAI's GPT-4o to reproduce the game. The returned code was entered into PICO-8, and feedback was given to the LLM. Such feedback contained information on missing mechanics, misrepresented graphics, balancing constraints, or errors in the returned code. Over the course of 30 queries, the basic game loop, as well as a main menu and a high-score screen, had been replicated. We show snapshots of the process in Fig. 25. The overall process took about 3 hours. While the final product looks similar to the original game, it still needs polish in balancing and mechanics.

This case raises several ethical considerations regarding the reproduction of an existing game. We doubt that the LLM had prior access to the *SuperHotRoids* game code; the reproduction was based solely on textual descriptions and screenshots provided by us. However, the resulting game is very much a derivative of the original. We therefore refrain from publishing the resulting game so as not to infringe on the game developers' copyright of *SuperHotRoids*. While this experiment demonstrates technical feasibility, similar methods could be misused to clone or imitate original works without consent, thus undermining the value of human creative labor.

3.8.3 Case 3: GenAI following a human creator

For Case 3, we followed a more traditional pipeline for co-creative tools: a human creator taking the initiative and the AI following [10]. We consider Case 3 a closer approximation of how GenAI will be used by indie developers, novices and students [4]. Specifically, the game was produced with entirely human game design decisions, and entirely AI-authored code. Using this process, we created a simple camera-based game (named *Webcam Island Builder* by the human author) connected to the "islands" theme of Case 1 (see Section 3.8.1). The LLM of choice was Google's Gemini 2.5 Flash via Cursor, and the game engine was Pygame (with Pymunk for physics).

The game concept revolves around webcam tracking: real-time footage of the player is used to physically interact with the game world. The player uses their hands, which act as rigid bodies in the game's physics, to guide randomly selected shapes onto a moving platform before time runs out. At the end, the player sees the "island" they created (see Fig. 26). The application is more of a toy, with no scoring or losing conditions.

[16] `https://superhotgame.com/`
[17] `https://en.wikipedia.org/wiki/Asteroids_(video_game)`

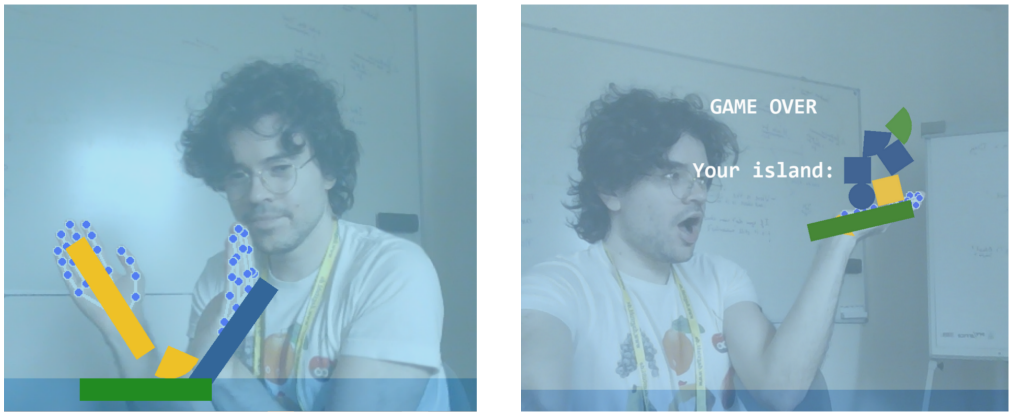


Figure 26 An in-progress (left) and end-game (right) screenshot of *Webcam Island Builder*.

This high-level concept was decided before engaging at all with GenAI systems. While first prompts were exploratory (e.g. to list appropriate libraries for game physics and webcam tracking), subsequent prompts were much more specific (e.g. "implement a timer that ends the game after 60 seconds"). A point was made to not allow creative decisions or design changes from the LLM. All prompts were designed to limit the LLMs' contributions to purely implementing clearly outlined functions.

The game was functional within an hour of this process, and only minor adjustments were needed after this. Overall, the game (while simple in terms of game logic) does what the designer expected and thus satisfies the – human – brief. Micro-adjustments were made after testing the game, on a minor scale (e.g. move speed of platforms), but they always followed the human designer's intuition and preferences.

3.8.4 Conclusions

Abiding by the intents of this working group, four games were created within a day[18], all with some degree of playability. We explored how three different LLMs handle game development tasks, and used three different game engines to make the games.

In terms of technical feasibility, PICO-8 seemed better suited for the LLM of choice (Claude) to design for, compared to Godot. We hypothesize on the reasons for this disparity in Section 3.8.1. More broadly, however, we can argue that lower-fidelity games with limited assets (graphics, audio, logic, interface, narrative) are easier for GenAI to produce. Such types of games are most often generated by a single indie developer over the course of a couple of days. This could suggest that GenAI can assist novices (who do not have a team or technical skills) to make a game faster, during a game jam. On the other hand, such indie developers are most vulnerable to GenAI misuse as "simple" games such as *SuperHotRoids* (see Section 3.8.2) can be easily reproduced causing Intellectual Property theft.

In terms of perceived creator agency, it was observed that engaging even superficially with the work-in-progress game and suggesting feedback to the LLM *did* increase the sense of ownership on the part of the human co-creator. However, we hypothesize that this sense of ownership is not how a game developer feels about their game; rather, it's closer to how a Quality Assurance tester (or fan) whose feedback has been heard feels about a product that is ultimately not theirs. Ethical issues of creativity, authorship [13], and intellectual

[18] Technically, each game took a couple of hours to make.

property (especially regarding Case 2) remain critical in this new era of GenAI. Moreover, it was surprising to find out that games made in a day seemed to us[19] to be "good enough". It is as exciting as it is worrying to speculate on what "good enough" will be perceived as in a year or a decade from now, and whether this shift will be due to a leap in Artificial Intelligence or due to a stagnation in human skill, appreciation and imagination [2].

References

1 Bartek Bogacki. Building a working game in an hour: My experience with vibe coding. `https://medium.com/agileinsider/building-a-working-game-in-a-single-\afternoon-my-experience-with-vibe-coding-8a5a02ddcabd`, 2025. Accessed August 2025.

2 Simon Colton. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Symposium on Creative Intelligent Systems*, 2008.

3 Michael Cook, Simon Colton, and Jeremy Gow. The ANGELINA videogame design system – Part I. *IEEE Transactions on Games*, 9(2):192–203, 2016.

4 Daniel Cox, John Murray, and Anastasia Salter. Routine, twisty, and queer: Pasts and futures of games programming pedagogy with no and low code tools. In *Proceedings of the 20th International Conference on the Foundations of Digital Games*, 2025.

5 Benj Edwards. Will the future of software development run on vibes? `https://arstechnica.com/ai/2025/03/is-vibe-coding-with-ai-gnarly-or-reckless-maybe-some-of-both/`, 2025. Accessed 21 Sep 2025.

6 Anna Jordanous. Four PPPPerspectives on computational creativity. In *Proceedings of the Second International Symposium on Computational Creativity*, 2015.

7 Anakaisa Kultima. Defining game jam. In *Proceedings of the Foundations of Digital Games Conference*, 2015.

8 Ichiro Lambe. The new surprising number of Steam games that use GenAI. `https://www.totallyhuman.io/blog/the-surprising-new-number-of-genai-games-on-steam`, 2025. Accessed August 2025.

9 Lexaloffle Games. PICO-8: A fantasy console. `https://www.lexaloffle.com/pico-8.php`, 2023. Accessed August 2025.

10 Antonios Liapis, Gillian Smith, and Noor Shaker. Mixed-initiative content creation. In Noor Shaker, Julian Togelius, and Mark J. Nelson, editors, *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*, pages 195–214. Springer, 2016.

11 Antonios Liapis, Georgios N. Yannakakis, Mark J. Nelson, Mike Preuss, and Rafael Bidarra. Orchestrating game generation. *IEEE Transactions on Games*, 11(1):48–68, 2019.

12 Juan Linietsky, Ariel Manzur, and the Godot community. Godot game engine. `https://godotengine.org/`, 2014. Accessed August 2025.

13 Jon McCormack, Toby Gifford, and Patrick Hutchings. Autonomy, authenticity, authorship and intention in computer generated art. In *Proceedings of the AAAI Symposium on Creative Intelligent Systems*, 2019.

14 Tim Merino, Sam Earle, Ryan Sudhakaran, Shyam Sudhakaran, and Julian Togelius. Making new connections: Llms as puzzle generators for the new york times' connections word game. *arXiv preprint arXiv:2407.11240*, 2024.

15 Noor Shaker, Julian Togelius, and Mark J. Nelson. *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*. Springer, 2016.

16 Tanya Short and Tarn Adams. *Procedural Generation in Game Design*. CRC Press, 2017.

17 Marvin Zammit, Antonios Liapis, and Georgios N. Yannakakis. CrawLLM: Theming games with Large Language Models. In *Proceedings of the IEEE Conference on Games*, 2024.

[19] We chose purposefully game development tools in which we were unskilled in.

3.9 Leveraging Jank

Timothy Merino (NYU – New York, US), Alena Denisova (University of York, GB), Antonios Liapis (University of Malta – Msida, MT), Adam M. Smith (University of California – Santa Cruz, US), and Yuqian Sun (Royal College of Art – London, GB)

3.9.1 Introduction

As Generative AI has surged in both popularity and cultural relevance, the various flaws of GenAI systems has also been dragged into the limelight. While various terms can describe these failures of generative systems, we adopt well known term from the gaming community, "Jank", to describe a certain subclass of erroneous outputs. In our working group, we first seek to examine what makes certain outputs entertaining, and then explore potential ways we can leverage these typically discarded janky creations in order to create something entertaining.

In the midst of widespread adoption of AI-generated content, as well as culutral opposition to "AI slop", there arises a form of nostalgia for the janky outputs of early image generation models. Legendary musician Brian Eno once said "Whatever you now find weird, ugly, uncomfortable and nasty about a new medium will surely become its signature. CD distortion, the jitteriness of digital video, the crap sound of 8-bit – all of these will be cherished and emulated as soon as they can be avoided"[1].

Our study of Jank serves to both categorize the emerging medium of "good Jank", as well as an attempt at intentionally leveraging it to create a game-like experience where Jank serves as the core mechanic.

3.9.2 Examples of Jank

Janky outputs exist in nearly every modality that Generative AI is applied to: text, image, video, etc. Funny examples are often widely spread on social media. For example, Apple's AI text summary feature made headlines due to it's humerous misinterpretations of text messages. For example, one summary reads "Multiple emergencies including house break-in, fire, and losing a Fornite match."

Google's AI integration with search has also produced notable and widely-shared fails. One search for "cheese won't stick to pizza" results in Google's LLM assistant suggesting the user mix some Elmer's glue into their cheese to add tackiness.

Jank is not exclusive to generative AI, and is a well known concept in video games, where it typically refers to frequently buggy game systems. Some games lean into janky game systems for comedic effect. A notable example is Goat Simulator, published in 2014 by Coffee Stain Studios. Described as a "chaotic sandbox", the buggy physics engine is the central component of the game loop, with players being encouraged to epxloid physics glitches to accomplish goals.

3.9.3 Defining Jank

AI Jank comes in as many forms as there are modalities for Generative AI, and we faced difficulty trying to assign a single canonical definition to the phenomenon. Our working group focused on identifying key properties of a generated output that Jank must have.

As a basis, a janky output is a (somehow) incorrect output of a system. The challenge comes in differentiating Jank from a typical error that may arise from insufficient training, lack of generalization, or a misconfigured system.

We identified five properties that help distinguish Jank from typical errors:

- **Unintended**: The viewer has some sense of what the intended, non-janky output of the system would have been like.
- **Unreproducible**: The system doesn't always produce Jank, it might do something like that again.
- **Inhuman**: It doesn't feel like you, as a human, could figure out how to practically reproduce the behavior even if it was your intention.
- **In-group specificity**: It is plausible that general audiences would not notice what's going wrong with the output.
- **Provocative**: The output is remarkable in some way, and evokes some emotional response in the viewer.

These properties aim to distinguish the type of Jank that may be fun and potentially useful, rather than simple failures of a generative system. When combining all of these features, we find that Jank first requires a capable and understandable generative model, where janky outputs are a subversion of the expected quality and subject matter of the typical output.

Because the janky behavior is unreproducible, you often experience it through a reliably captured recording of the original behavior. If the original intent is not obvious, people may not perceive it as a satisfactory Jank. Once we find ways to humanly reproduce it on purpose, even the original example of the Jank becomes less remarkable.

3.9.4 Project

We attempted to further explore how Jank can be leveraged for entertainment, creating a humorous look at the failures of text-to-image generation.

We first captured real world images of the rooms at Dagstuhl, as well as an image of the floorplan. Our goal is to create an alternate "Jank-dimension" version of the space we occupied.

First, we utilize the "recursive ChatGPT image" method to introduce spelling errors and layour errors in the image of the floorplan. After 10 iterations, we obtained some interesting mis-spellings of existing rooms to serve as the map for our exploration game.

Then, we tried to match the hallucinated room titles to their original rooms, selecting a picture to use as the basis for each room. We then used Midjourney's image editing feature to generate Jank variations of each image based on the room name.

We chose Midjourney as our image model because of the optional "chaos" parameter they expose in the interface, a scalar value that can be set when generating or editing an image. This parameter "lets you add more variety to the image results you get from each prompt", though they warn "higher values mean the images can be quite different and may not stick as closely to your prompt, giving you unpredictable results". We find that this approximately serves as a Jank parameter, with higher values often resulting in bizarre and hilarious results.

For each room, we masked out certain regions of the source image to replace using the model, and provided the text prompt "A group of researchers at the New Frontiers in AI for Game Design seminar at Schloss Dagstuhl. A big sign on the wall says "<room name>"". We set the chaos parameter to 100 each time, and continued generation until we got an image met our Provocative criteria.

3.9.5 Conclusion

We have identified five key properties that distinguish entertaining "jank" from simple errors in generative AI: unintended, unreproducible, inhuman, in-group specific, and provocative outputs. Our experimental project – creating a jank-dimension Dagstuhl using Midjourney's

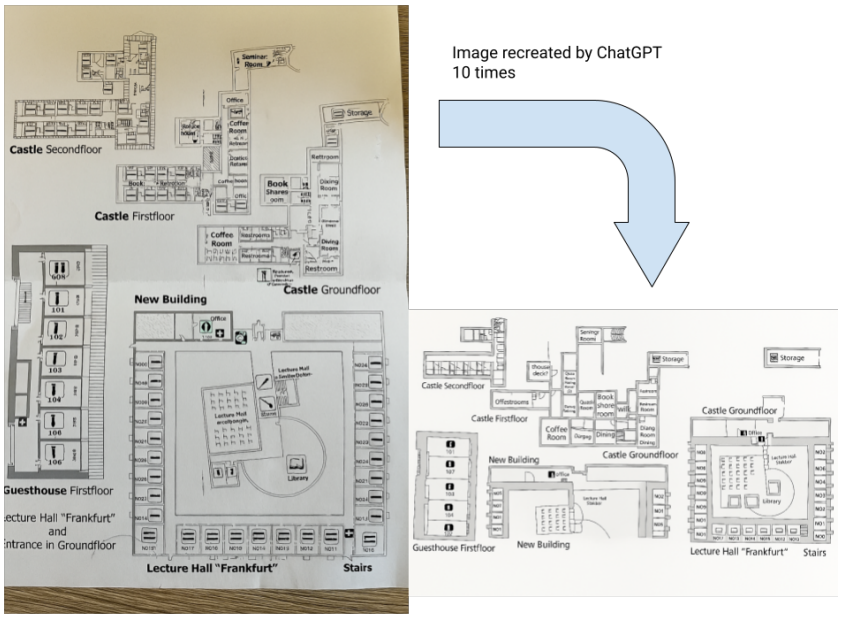


Figure 27 The original floor plan and final hallucinated floorplan produced by ChatGPT.

Figure 28 Process for generating a final image included in our game. Infilling operation was repeated until we produced a sufficiently striking output.

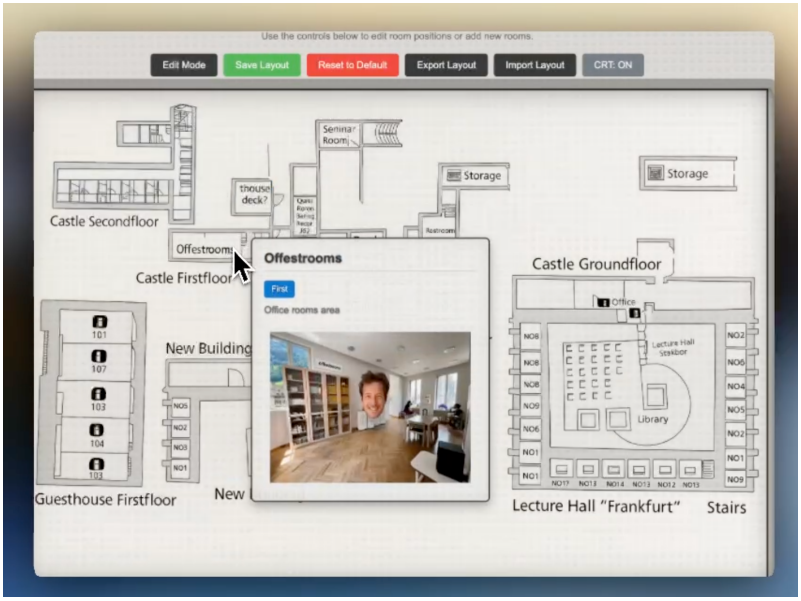


Figure 29 Interactive map of Dugstughl with jank output.

chaos parameter – demonstrates that these typically discarded outputs can be leveraged as a creative resource. Rather than viewing jank as failure, we propose embracing it as an emerging aesthetic that, as Eno predicted, may become a cherished signature of early generative AI. Future work could explore tools specifically designed to produce controlled jank or investigate how audiences' perception of these artifacts evolves as AI systems mature.

References

1 Brian Eno. *A year with swollen appendices: Brian Eno's Diary*. Faber & Faber, 2020

3.10 Handmade Blaseball

Younès Rabii (Queen Mary University of London, GB), Claus Aranha (University of Tsukuba, JP), Brian Bucklew (Freehold Games – Walkerton, US), Michael Cook (King's College London, GB), Rémy Devaux (Punkcake Délicieux – Cenon, FR), Matthew J. Guzdial (University of Alberta – Edmonton, CA), Florence Smith Nicholls (Queen Mary University of London, GB), and Yuqian Sun (Royal College of Art – London, GB)

This Dagstuhl report details the results of a session entitled "Handmade Blaseball". The focus of the session was on discussing analogue procedural content generation. This session's members were Brian Bucklew, Claus Aranha, Florence Smith Nicholls, Matthew Guzdial, Michael Cook, Rémy Devaux, Younès Rabii, and Yuqian Sun.

3.10.1 Framing

Our initial motivation for this working group was to explore how we could recreate the process of Automated Game Design without involving any computers. Many of us were motivated by a sense of disillusionment with the trajectory of games AI, and analogue prototyping offered

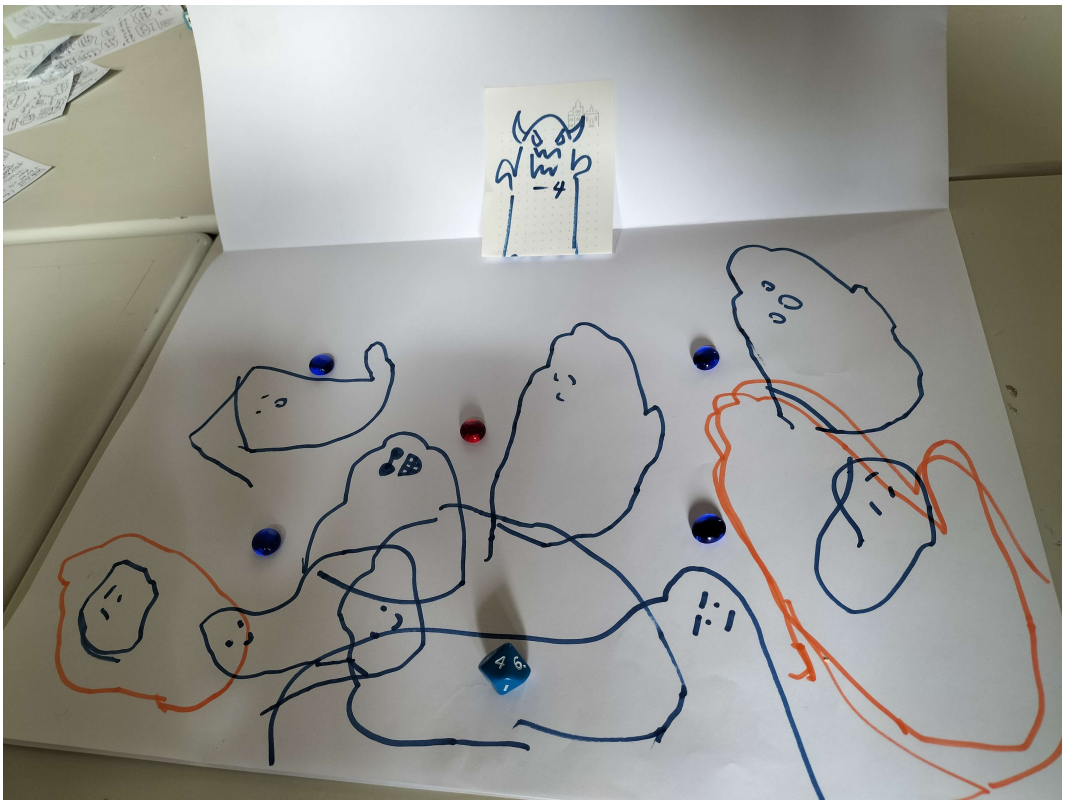


Figure 30 Picture from a game of "*A Monster Haunted By 1000 Artists*", itself generated by playing "*Blank-Page-Boogie-Woogie*" [4].

a way to foster surprise, emergence, and playfulness outside of computational constraints. Our guiding reference was *Blaseball*: a cryptic simulation video game which left a lot of room for audience participation and, like an improvised performance, constantly felt like it could truly go anywhere. Some of the reasons why *Blaseball* had to be brought to an end were the increasing logistical and technical costs of maintaining this setup at that scale. The question we set to investigate was: What can we do in that design space, at a local scale and with little to no computation?

3.10.2 Process

Not all members of the session were familiar with Blaseball and similar community-driven games [1]. As such, we began by overviewing the area to ensure all members of the discussion were on the same page. Due in no small part to the makeup of the group, the conversation quickly shifted to automated game design and how one could accomplish this in an analogue (i.e., non-digital) fashion.

We quickly proposed an initial version with very simple rules, namely:

- Assumptions: board, pieces, turns
- Step 0: Lines, veils, wishes for game design
- Step 1: Come up with number of players collaboratively
- Step 2: Assign 1+ to each player:
  - Board gen rules (give 1 example board)

- Piece gen rules (give 1 example piece)
- Piece defining rules (give n example teams)
- Piece placement on board rules (give 1 example board with pieces)

- Step 3: Generating your game.
- Step 4: Assign 1+ to each remaining player:
  - Win conditions
  - Turn start
  - Turn end
  - Arbitration

We nicknamed this "Four G's" or "Four Gees" as it was a Game Generator Generator Game. The idea was that some parts of the game would happen in secret (Step 2) but each person would give some information (a specific output from their generative rules) to allow for cohesion. We played an initial narrative-driven chess-like game that we collaboratively created and then reflected on this process.

After playing the chess-like game, we reflected on the experience, with each member having different takeaways. Some members wanted more iterations on the design, others felt that iteration was a trap. Some wanted more simplicity, others wanted more chaos. As such, we split up and each member created their own variants independently.

3.10.3 Results

We lack the space to fully overview all of the seven game creation variants. Many were simple variations on the initial game described above, such as adding more explicit instructions [2] naming every game component or picking an explicit theme.

We focus our attention on three variants that were more strenuously tested at Dagstuhl, Michael Cook's "Blank Page Boogie-Woogie" [4], Claus Aranha's "Variation 5" [3] and Rémy Devaux's unnamed variant.

For "Blank Page Boogie-Woogie" the aim was to develop a process that felt very easy to follow, almost like a folk game that you could describe to someone verbally and pass on that way too. The game required the use of pens and paper, along with the optional inclusion of other objects. The game included an explicit 'fix the design' phase in order to lessen the concerns of producing something playable initially. Based on the playtesting this seemed very effective.

"Variation 5" was more explicitly inspired by the author's prior experience with Role Playing games and Board Games. Thus it ended up with explicit turn taking rules, and explicit separation of responsibility between the players, to try to make sure that all players have a fair chance to contribute to the design, even if they are unfamiliar with this kind of exercise. It also included a more abstract rule about "naming" the rules previously described, both to add a touch of whimsy to the process, as well as to allow a chance to review the final design.

In Rémy's variant, all assumptions in the game were chosen collaboratively, with rules designed simultaneously due to card draws. Finally at the end, the game permitted two changes to the design as an explicit reflection.

After presenting all seven games to one another, we selected the three variants above to playtest, using Variation 5 twice and the others once to create a total of four games. With Variation 5 we produced "Fruits", a physical game about getting a piece of fruit as high as possible and "Heaven or Hell", a game about playing out Christian theology. With Blank Page Boogie-Woogie we created an unnamed game about rolling dice and drawing ghosts. With Rémy's variant we created an analogue game simulating animals escaped from a zoo.

3.10.4 Observations

We found a number of interesting observations when discussing the process of designing the game generators, creating the games, and playing the games. First was the importance of differentiating what we meant in terms of the different roles that people took on in these experiences, whether we meant a designer of a generator, a designer of a game, or a player of a game. We opted to use the term "performer" in several instances due to the ambiguity of "player" in this context.

Second, across these game variations we had a spectrum of information sharing between participants in the game generation process. When participants did not have access to the other designer's rules, they often trended towards making "game agnostic" rules that could have functioned in any game. In comparison, when rule creation was public, individuals could more easily build on top of one another's design.

Third, an effect we dubbed the "Ouija Board" effect during the session, but is more closely related to the notion of sensemaking from psychology [5]. In this case we repeatedly found ourselves, apparently by happenstance, having designed a game that had interesting things going on in it, despite the fact that the components of that game were designed independently. We called this the "Ouija Board" effect as it felt as though we had subconsciously manipulated the game together to produce a coherent outcome.

3.10.5 Reflections

In reflection on this session we found that there was definitely something compelling about the process of designing these analogue game generators, designing the specific games collaboratively and performing them in front of an audience. We feel that there are likely connections here to improv, informal game design, dadaist games such as *exquisite corpse* and folk games. We believe there is potential in encouraging the same group to continue iterating on the same generator over multiple sessions, allowing them to build their own meta-narrative and leaving room for a communal emergent narrative. We think these game generators may also be useful vectors for studying game design processes, and hope that future researchers can more fully investigate this possibility.

References

1 Sam Rosenthal. "Welcome to Blaseball." Blaseball, accessed August 12 (2021).
2 Matthew Guzdial. "Game Generator Generator Games." `https://mguzdial.itch.io/game-generator-generator-game`, Accessed September 27 (2025).
3 Claus Aranha. "Variation 5" `https://caranha.itch.io/variation5`, Accessed September 27 (2025).
4 Michael Cook. "Blank Page Boogie-Woogie" `https://illomens.itch.io/blank-page-boogie-woogie`, Accessed September 27 (2025).
5 Helms Mills, Jean, Amy Thurlow, and Albert J. Mills. "Making sense of sensemaking: the critical sensemaking approach." Qualitative research in organizations and management: An international journal 5.2 (2010): 182-195.

3.11 "But What About A Secret Third Thing": Exploring Playful Transgressions In Video Games

Dipika Rajesh (University of California at Santa Cruz, US), Brian Bucklew (Freehold Games – Walkerton, US), Younès Rabii (Queen Mary University of London, GB), M Charity (University of Richmond, US), and Adam M. Smith (University of California – Santa Cruz, US)

When considering the act of playing games, much of the research and design discourse tends to center on canonical forms of play that are aligned with a game's intended mechanics, rules, or narrative goals. Yet, alternative and subversive forms of engagement such as modding, speedrunning, or adapting "playground rules" represent an underexplored but rich avenue for both AI research and games research more broadly. These practices not only challenge the boundaries of how games are designed to be played but also open new opportunities for computational systems to understand, support, and even generate diverse playstyles.

Modding has long been recognized as an important approach to game development through the community creation of custom extensions and modifications to existing game software. Scacchi [1] emphasizes how modding serves as a form of extension of software, allowing for the adaptation and transformation of game systems beyond their original design. Despite this, there remains relatively little literature in AI or technical research that systematically maps these practices or explores computational systems designed to support them. Recent developments indicate a growing interest in this area: NVIDIA's RTX Remix platform [2] enables modders to enhance classic games with modern graphics, including AI texture enhancements and ray tracing. However, these advancements are still emerging, and the broader field remains underexplored, particularly in terms of systematic frameworks, user-friendly tools, and community-driven AI applications for modding that integrate playstyles beyond the developer's intent.

The motivation behind this working group was to better understand the current landscape of subversive playstyles and to ideate ways of supporting them through new tools, frameworks, and collaborations.

3.11.1 The Ontology of Playful Transgressions

During the first part of the morning, the working group focused on surveying the different types of gameplay that can be considered playfully transgressive. This initial mapping produced an expansive list that included speedrunning, ROM hacking, playground rules, modding, game corruptions, the use of cheat codes, and save scumming. Examining these diverse playstyles highlighted the many ways in which players depart from the conventional play patterns anticipated by a game's design. This discussion immediately raised a critical question: what, precisely, is being subverted or transgressed through these practices? We identified that beyond the formal rules of a game, certain playstyles also transgress its coded affordances, as well as the legal, moral, and even political frameworks that games may inherently embed. Developing this landscape offered a clearer view of the vast potential for new systems and methods to support such unconventional modes of play.

3.11.2 "I Make The Rules Around Here!": A Theory of Playful Transgressions

We developed a theoretical framework for defining playful transgressions by identifying three key dimensions: Players, Designers, and the Computational Medium. Within this framework, we defined normal play as activity that falls within the intended or allowed range across all three dimensions. Considering only the Player and Designer dimensions, there exist experiences that are acceptable and desirable to both but constrained by technical limitations, such as bandwidth or storage; these are not considered transgressive play. In contrast, when focusing on the Player and Medium dimensions while excluding the Designer, leveraging the medium to create exo-designed experiences, such as mods, hacks, or house rules, constitutes transgressive play. Finally, considering the Designer and Medium dimensions without accounting for the Player highlights practices that manipulate or restrict the player, such as inserting ads or extracting private data. We also stretched the framework by looking at each dimension on its own. To make these ideas more tangible, we created a zine that illustrated the theory and shared it with seminar attendees.

3.11.3 A Game Designers Guide for Enabling Playful Transgressions

We approached the space of playful transgressions from a game designer's perspective and mapped out ways in which designers could intentionally structure their games and systems to support versatile playstyles. From this exploration, we identified five key considerations and highlighted examples of systems that embody each principle: choice of platform (e.g., Skyrim on PC vs. Xbox), modular and data-driven system design (e.g., Caves of Qud), explicit mod hooks (e.g., Minecraft), documenting the design process (e.g., Balatro), and observing, learning from, and supporting the community (e.g., Stardew Valley). To disseminate these ideas, we designed a zine that encapsulated these concepts and distributed it to the attendees of the remainder of the seminar at the conclusion of the workshop.

Figure 31 Game Genie, an example of an accessible system for playful modification.

3.11.4 Designing a Prototype System for Supporting ROM Hacks

Game modding and ROM hacking present themselves as spaces where players require substantial programming expertise, as well as significant time or resources to outsource technical work. These barriers restrict access to the creative potential of modifying and reimagining games. In our discussions, we drew inspiration from two important sources: the Game Genie [3] and Blockly [4] (Figure 31). The Game Genie [3], a hardware device popular

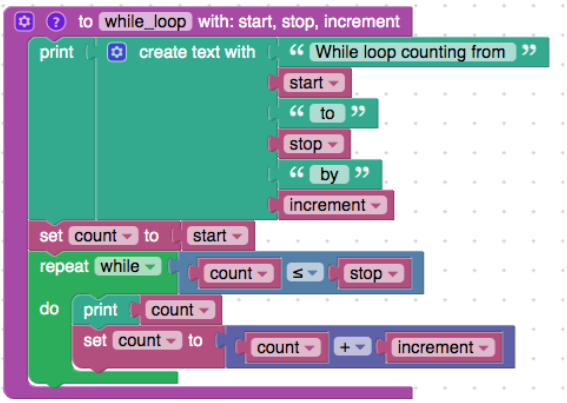


Figure 32 Blockly, an example of an accessible system for playful modification.

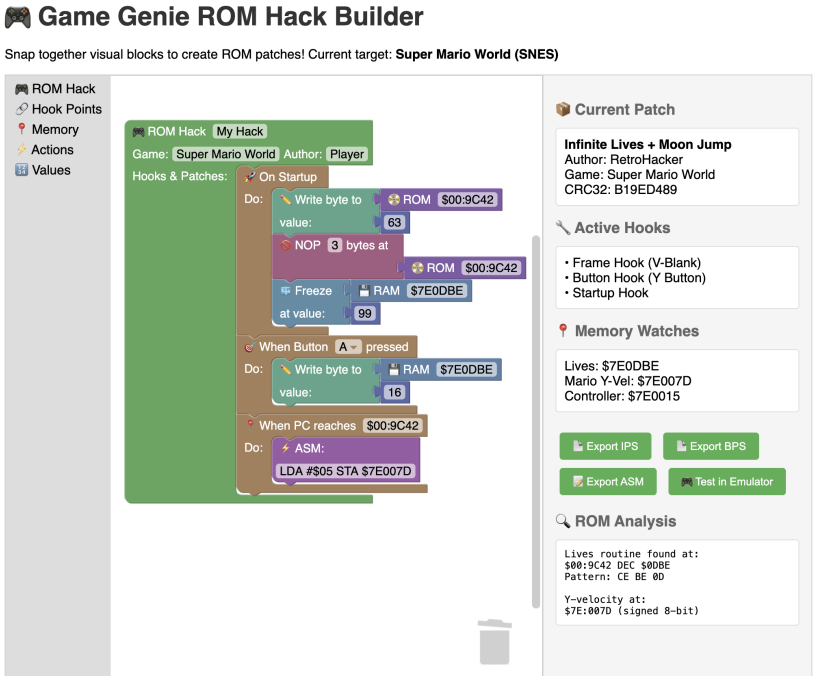


Figure 33 Prototype user interface for supporting ROM hacks, ideated with GenerativeAI to combine accessibility and visual programming concepts.

in the 1990s, exemplified how cheat codes and memory manipulation could offer everyday players accessible ways to alter a game's behavior without requiring direct interaction with its source code. It made the act of "rewriting" a game approachable, playful, and within reach of a broad audience. Blockly [4], in contrast, demonstrates the power of visual, block-based programming to lower the threshold of entry into computational thinking. Its drag-and-drop interface to construct logic and behavior in intuitive, modular ways. Additionally, many ROMs are thoroughly mapped by fan communities (for example, see [5]), whose documentation makes hidden game structures easier to interpret. These mappings could be valuable for creating accessible tools, since they transform opaque technical details into interpretable knowledge. By combining the accessible nature of Game Genie, the visual

programming paradigms of Blockly, and the ROM mapping provided by the community, this UI (Figure 33) design and the potential system powering it could make modding approachable to a wider range of players.

3.11.5 Conclusion and Future Work

This working group highlights the considerable potential for creating more accessible systems that enable a wide range of playful and transgressive gameplay. Future work could involve developing a full prototype of the GUI system for ROM hacking in a single game, or designing an AI-powered tool capable of supporting multiple games and playstyles. Beyond tool creation, there remains substantial research to be done in understanding how players engage in these unconventional forms of play and the rich communities that emerge around them, such as those on Twitch and other social platforms. Exploring these communities and practices more deeply will provide valuable insights into the ways games can be expanded and experienced across a variety of playstyles.

References

1 Scacchi, W. (2011, October). Modding as an open source approach to extending computer game systems. In IFIP International Conference on Open Source Systems (pp. 62-74). Berlin, Heidelberg: Springer Berlin Heidelberg.
2 Usmani, N. (2025, March 13). Breathing New Life Into Games With RTX Remix. NVIDIA Blog. https://blogs.nvidia.com/blog/rtx-ai-garage-rtx-remix/
3 NES World. *Game Genie – The video game enhancer*. Retrieved from https://www.nesworld.com/
4 N. Fraser, "Ten things we've learned from Blockly," in 2015 IEEE Blocks and Beyond Workshop (Blocks and Beyond), 2015, pp. 49–50.
5 Data Crystal, "The Legend of Zelda ROM map." Available: https://datacrystal.tcrf.net/wiki/The_Legend_of_Zelda/ROM_map
6 Nathalie Lawhead. *Electric Zine Maker*. https://alienmelon.itch.io/electric-zine-maker

3.12 PCG for Keepsake Games

Florence Smith Nicholls (Queen Mary University of London, GB), June Bhartia (Télécom Paris, FR), Michael Cook (King's College London, GB), Younès Rabii (Queen Mary University of London, GB), Dipika Rajesh (University of California at Santa Cruz, US), Anne Sullivan (York University – Toronto, CA), Yuqian Sun (Royal College of Art – London, GB), Nicolaas Vas (Billund, DK), and Sabine Wieluch (Universität Ulm, DE)

The term "keepsake game" was coined by Shing Yin Khor in 2021 to refer to "games that produce beautiful, memorable artifacts, through the process of playing the game" [1]. Keepsake games involve creating or modifying an original artifact as part of the gameplay process, guided by prompts as part of the design. Many keepsake games, such as Jeeyon Shim and Shing Yin Khor's *Field Guide to Memory*[20], are analogue games. In this Dagstuhl working group, we set out to explore the potential of designing keepsake games that included

[20] https://jeeyonshim.itch.io/field-guide-to-memory

a digital procedural system to prompt the creation of an analogue keepsake. We discussed a typology of keepsake games, complementary procedural systems and design constraints, before splitting into four subgroups to create prototypes. Through prototyping, members of the working group also explored analogue procedural systems and digital keepsakes, further challenging any strict demarcation in this hybrid design space.

3.12.1 Motivation

There has been limited explicit academic engagement with Khor's concept of keepsake games. One clear outlier to this is a chapter in *The Routledge Handbook of Role-Playing Game Studies* within the context of text-based role-playing games [2]. Though the term "keepsake game" was coined recently, arguably games that could fall into this category have a longer history. A 1997 paper published in the *Proceedings of the International Conference on Cognitive Technology* [3], for example, discusses digital augmentation of "keepsake objects." Sullivan et al's *Loominary* captured narrative choices in a digital Twine game by having players use a loom as a controller, thus producing a physical artefact as part of the gameplay process [4].

Loominary's hybrid digital/analogue interface was a key inspiration for the session. The first author was interested in exploring the possibility space of such hybrid keepsake games, especially in conjunction with procedural systems. Procedural content generation (hereafter PCG) is, broadly speaking, the generation of content algorithmically. Given that keepsake games often include human-authored algorithms for play through creative prompts, it was theorised that incorporating PCG systems would be complementary. Furthermore, we were interested in exploring how the uncanny poetics and texture of PCG [5] might contribute to keepsake game design.

The first author chose the "no generative AI content" category for this working group.

3.12.2 Designing Procedural Keepsakes

During the first part of the session, we discussed several different aspects of keepsake game design as a group. This discussion is summarised in the three subsections below.

- ***Keepsake Typology***. In terms of different types of keepsake, we came up with the most examples under the sub-category of "crafted items." Crafted items are material agnostic, the emphasis is more on the process of physically creating a piece, such as paper weaving and even edible crafts. Crafted keepsakes are complimentary to the idea of keepsakes as a gift, which would be broadly applicable to other personalised keepsakes, such as postcards. We also discussed digital keepsakes, such as character creators and USB drives.
- ***Complementary Procedural Systems***. We discussed both digital and analogue procedural systems that could be incorporated into keepsake games. In terms of the former, existing digital games with PCG such as *Minecraft* were suggested. Another system was the *Tracery* text generation tool. We also had suggestions for analogue PCG, such as dice rolls, Tarot cards and the cut-up method.
- ***Design Constraints***. Potential design constraints for keepsake games fell mainly under the following areas; theme, material, duration, number of players, location and familiarity. Familiarity was particularly applicable in terms of the aforementioned crafting techniques, as potential players may need to learn a new skill as part of crafting a keepsake. In addition, players might need access to specific craft materials. Though we did not extensively explore this in our discussion, it is important to point out that accessibility is thus a key concern in terms of keepsake design.

3.12.3 Prototypes

In the second part of the session, we split into four subgroups to rapidly prototype keepsake games. Each of these prototypes, and the design considerations that went into them, are summarised below.

3.12.4 A Sending

A Sending [21] is a postcard keepsake game (the name riffs on Shing Yin Khor's keepsake game *A Mending*[22]). Players use an online village generator tool to create the map of a fictional village, are instructed to design a stamp based on it, and write an accompanying postcard with details of their trip to this fantasy place.

Role of PCG. We decided to use an existing digital map generator tool for ease of prototyping. This is watabou's *Medieval Fantasy City Generator*[23], freely available on the itch.io platform, and the developer has stated they are happy for images generated from it to be used in other creative works. Furthermore, the generator can be used in-browser, so it is fairly easy to access.

Design Constraints. One of the major design constraints was writing the keepsake game with the expressive range of the *Medieval Fantasy City Generator* in mind. The generator only produces schematic details, as opposed to a large range of distinct, annotated building types, so it was important to keep in mind what would be present in any given map and so would complement any written prompts. In addition, the instructional part of the keepsake game itself was constrained to an 8-page zine.

Reflections on the Design Process. The zine constraint was useful as it encouraged us to write efficiently, while still allowing a little room for flavour text. Using watabou's generator was a great starting point for thinking about a game where you are essentially roleplaying as a tourist in a village based on a procedural map. However, making a bespoke generator would have allowed for a more sophisticated relationship between the analogue keepsake creation process and the digital map. We encountered some issues attempting to repurpose watabou's generator for our needs, partly due to the constraints of modern browsers, all of which are easier to mitigate if building a generator from the ground up. Furthermore, we would have benefitted from a more iterative process in which we had the opportunity to play test the game ourselves.

3.12.5 A – Zine

A-Zine (pronounced "A-to-Zine") is a collection of 26 pocket zines. Each zine has a front-page with a fixed title, an empty square and a field to write the author's name. Each zine is empty, containing only two words written in red, in different pages. Red words are hyperlinks leading to the zine with that same title. Readers are invited to pick the zine they like, read its short story and follow hyperlinks to other zines. If the zine they opened is empty, they are invited to fill it using its title as an inspiration. They're invited to include the fixed red words within its content – as part of a text, for example.

[21] `https://florencesmithnicholls.itch.io/a-sending`
[22] `https://sawdustbear.itch.io/a-mending`
[23] https://watabou.itch.io/medieval-fantasy-city-generator

Figure 34 An excerpt of a map produced by watabou's Medieval Fantasy City Generator.

Each zine has a title starting with a different letter of the English alphabet. The first one, titled **Atlas**, explains how to read and fill the zines.

Role of PCG. PCG wasn't used for this first prototype, but the system was designed to easily include it in the future. There are two components for which a PCG system would be relevant: picking names for the zines –which often correspond to locations– and generating the "map" i.e. the graph of connections between zines.

Figure 35 The graph of connections between zines (top right) and a selection of zines for corresponding locations

Design Constraints. Our initial goal was to develop a system akin to a game engine that would support player creativity. Our core examples were platforms such as Twine, Bitsy and Downpour. Development time was limited by the working group's duration of one day. We quickly settled on two self-imposed constraints that were our design ethos:

1. We wanted players to create zines and
2. We wanted as little friction as possible between the act of stumbling upon the game and being able to play it.

Reflections on the Design Process. Our initial discussions circled around the idea of letting players create small dungeons, puzzles or escape rooms within a zine. Subsequent players would have to solve the zines by entering a password on a website, which would lead them to another zine. When discussing the infrastructural technology choices to enable this, we realised that it would create a lot of friction: scanning a QR code, needing a phone, needing an internet connection, entering a password, creating an entry on a website, etc. All those actions clashed with our accessibility goal. In the end, we made the choice of removing smartphones and computers as much as possible from our system.

This principle quickly led us to make practical decisions adapted to a physical medium. We needed the zines to be self-contained yet connected, we needed them to be easy to understand on a first glance and offer vast possibilities of self-expression. We ultimately converged on designing a small library of nearly empty but interconnected zine templates.

3.12.6 Color Collector

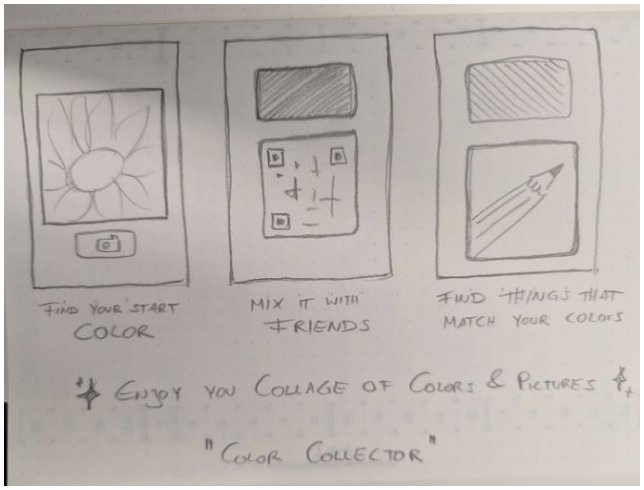


Figure 36 A sketch of the Color Collector interface

Color Collector is a collage and connection game played with a smartphone. In the beginning, each participant is prompted to photograph an object that has a color they like. This color is extracted from the image and becomes their starter color. From then on, participants can either:

After the event, each participant has created their very own collage of little moments and hopefully has found new friends or connections during the color mixing process.

Role of PCG. There is no software side of PCG – the participants are in a way the content generators.

Design Constraints. The goal of this keepsake game was bringing people together (in a setting like a conference or another small event) and giving them a purpose to interact with each other. Another goal was to mix physical and digital interactions. It was important to us that both the physical and digital part of the keepsake game were easily accessible and understandable.

Reflections on the Design Process. The prior mentioned constraint quickly narrowed the interaction down to a digital interaction via smartphone. We found it very helpful that so many smartphone features, such as the camera, can be accessed via the browser which allowed us to design the game without needing a stand-alone app, which makes it a lot easier to access.

3.12.7 Postcards for Dagstuhl

Postcards for Dagstuhl is a physical game dedicated to the seminars that are held at Dagstuhl, although it could easily be adapted for other gathering spaces. The output of the game is a hand-decorated postcard which reflects something about the player and their time at Dagstuhl. Inspired by instructional art, during the game the player is given three sets of prompts to choose from with instructions on how to decorate their postcard. The first table has prompts which reflect the player's own research or personality, the second table has prompts about the player's experience at Dagstuhl, and the final table has prompts about who to give the postcard to when they are done, if they don't choose to keep it for themselves.

Role of PCG. As a physical game, computational PCG did not play a part in the process. However, players could use dice in choosing the prompts they responded to, which added some simplified, traditional aspects of PCG.

Design Constraints. The design was constrained based on wanting the game to be approachable by non-artists and non-gamers, as well as being able to realistically fit into a typical Dagstuhl Seminar experience. To keep the barriers to entry low regardless of a player's artistic or gaming background, the game rules suggested different options for decorating the postcard beyond traditional forms of art, such as writing, collage, and making graphs. Additionally, the game is primarily solo play, and without competition to make it less intimidating for non-gamers. To help the experience fit into a busy Dagstuhl schedule, the game was made so all the instructions could fit into a one-page zine, and it was designed to take 15-30 minutes.

The prompts for the game were all shaped by these constraints and the goals of the game to deepen community building and encourage reflection. The first set of prompts asks the player to represent themselves or their research, the second to reflect on their time at Dagstuhl, and the third gives the player choices for whom to give their postcard to.

Figure 37 A series of postcards made at Dagstuhl Seminar 25292, referencing other working groups.

Reflections on the Design Process. The game went through several iterations, with the prompts getting refined to better fit the design goals. However, the most successful part of the game was the format of the keepsake: postcards.

Using postcards was particularly effective because it provides a small canvas, which is less intimidating than a full piece of paper. It also takes less time to decorate which keeps the game experience shorter. Postcards are also already associated with less serious forms

of writing in American culture, and they are also seen as something you give to someone else. These associations fit particularly well with the design goals of the game. Despite the experience not generally taking much time, the hand-made and physical nature still gave players some time to reflect on their experiences. And more importantly, the physical, hand-made nature of the keepsake generally made them feel more special when they were given to someone else.

3.12.8 Conclusion

Four prototype keepsake games were made as part of this working group. While the initial prompt was to create physical keepsakes games that used digital PCG systems in some way, many of the prototypes ended up exploring forms of analogue PCG [6]. Overall, there was a concern for accessibility in terms of both artistic and gaming experience and availability of specific digital platforms. We believe there is great potential for further exploring and experimenting with PCG keepsake games, especially in terms of creating bespoke generators for this purpose.

References

1 Shing Yin Khor. *on keepsake and connected path games*. Patreon. Online: https://www.patreon.com/posts/on-keepsake-and-47599952, 2021
2 Jessica Hammer and Paul Czege. "Text-Based Role-Playing Games." In *The Routledge Handbook of Role-Playing Game Studies*, pp. 171-184. Routledge, 2024
3 J. W. Glos and J. Cassell. "Rosebud: a place for interaction between memory, story, and self." In *Proceedings of the 2nd International Conference on Cognitive Technology*, IEEE Computer Society, USA, 88, 1997
4 Anne Sullivan, Joshua Allen McCoy, Sarah Hendricks, and Brittany Williams. "Loominary: crafting tangible artifacts from player narrative." In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 443-450. 2018.
5 Isaac Karth. "Preliminary Poetics of Procedural Generation in Games." In *Transactions of the Digital Games Research Association* , 2019.
6 Gillian Smith. "An Analog History of Procedural Content Generation." In *Foundations of Digital Games Conference*, 2015.

3.13 Dagstyle

Nicolaas Vas (Billund, DK)

This report details the development of Dagstyle, a visual system created in the months leading up to Dagstuhl Seminar 25292, intended as an inspiration and development tool for attendees of the seminar.

3.13.1 Background

Visual systems are a defined set of rules for creating consistent assets, layouts and designs. They are often used by companies, brands and products to be easily recognized and remembered in the minds of consumers. Visual systems can be thought of as individual components, which are arranged into assemblies, which are finally placed into applications.

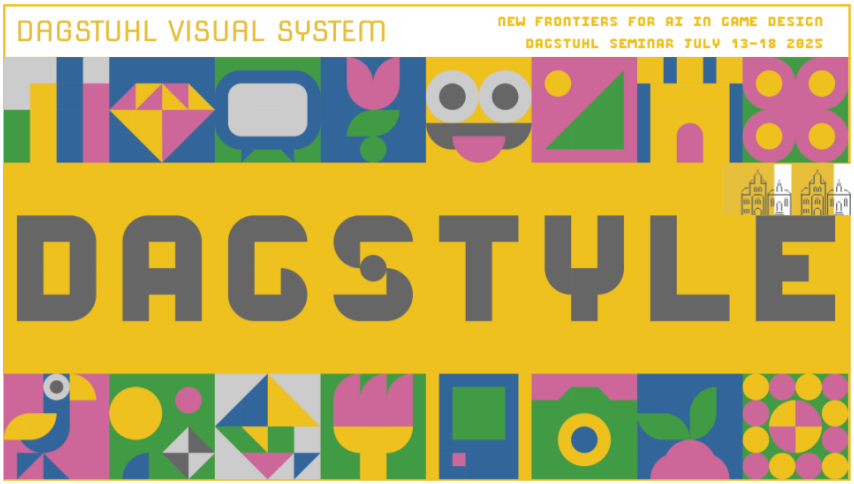


Figure 38 A title slide from the introductory talk presenting Dagstyle, made with the visual language.

Benefits of visual systems include their ease of use, their flexibility to scale across media and audiences, and a tendency for their constraints to cultivate creativity. In a time when Gen AI disrupts the status quo in the pursuit of high fidelity, simple limited systems can also provide a space for all to experiment.

In addition to brand design, visual systems can be found in many places, such as national flags, the pixel art of classic video games, tangrams, Nintendo Miis and even the brickwork in Victorian townhouses. Dr. Martin Lorenz advocates for a move away from logos and towards flexible visual systems that can be easily scaled across different audiences and in different media.

In preparation for the seminar, the author was inspired by the visual language of DROPS[24] (The Dagstuhl Research Online Publication Server), which is optimized for text and icons with a pixel and line drawing style. A limited colour palette of four colours allows for clear contrast between white, grey and yellow, with blue for hyperlinks. Could this be extended to allow for a wider expressive range, to create a simple and accessible visual system for the seminar attendees to use and further develop during the week?

3.13.2 Development

Early exploration pushed the limits of the DROPS style, and endeavoured to define the shapes, colours and grid rules that form the foundation of Dagstyle.

Shapes – The allowable shapes were expanded from pixels to include squares, circles, triangles and quadrants.

Colours – The allowable colours were expanded from four to six; White, Grey, Yellow, Blue, Pink and Green. While the four colour map theorem states that only four colours are needed to colour regions so that no adjacent regions share the same colour, the addition of Pink and Green introduce new possibilities to represent nature and more colourful ideas.

Grid – All Dagstyleimages should be limited to placement on 8x8 or 16x16 grids, with the option for layering and shapes to be rotated by 90 degree increments. Assemblies can also be used as Components, allowing for fractal details.

[24] https://drops.dagstuhl.de/

After defining these simple rules, it was then possible to play with the system to create patterns, fonts, flags and 100 interesting things.

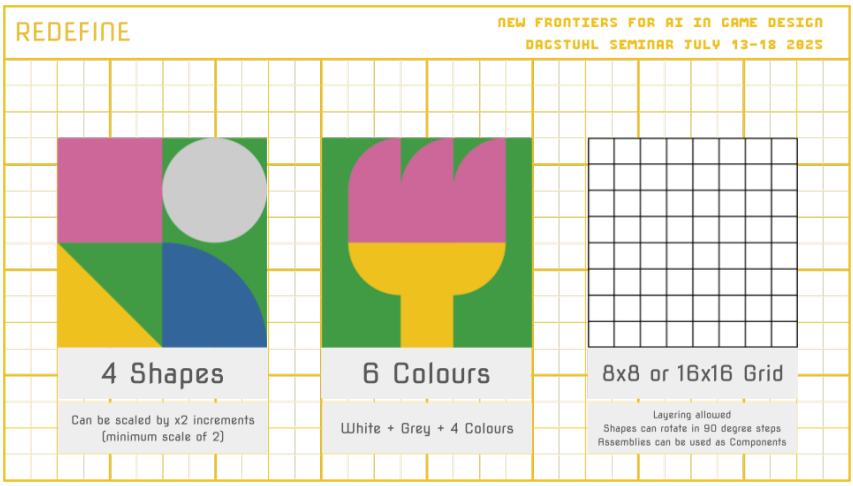


Figure 39 A visual representation of some of Dagstyle's defining features.

3.13.3 Result

The Dagstyle visual system consisted of the following components, which were provided to seminar attendees as inspiration and use during the seminar.

Dagstamp – A 2D Building system of icons and assets, that could be used in presentations and prototypes. Assembly sheets for hundreds of individual assets were provided, in addition to PNG and SVG images.

Dagslide – A Google Slides presentation template that could be used for sharing the results of workgroups throughout the week.

Dagscript – An OpenType font and writing system.

Dagzine – A self-publication zine template to introduce zine making.

Daggame – An invitation was given to seminar attendees to consider how the Dagstyle visual system and principles could be extended into an interactive video game format.

3.13.4 Name Tag Building System

With the hopes of providing a fun icebreaker activity, assets from Dagstamp and Dagscript were curated into an array of name tag building components. When supplemented by scissors, markers, glue and thermal photo printers, these symbol symbols were transformed by seminar attendees into a dazzling array of colourful name badges. The author was thrilled by the collective creativity and how well everyone worked together, many of whom were meeting each other and attending this seminar for the first time.

3.13.5 Conclusion

The Dagstyle visual system was well received, and served its purpose of inspiring a playful prototyping approach during the seminar. It was used, extended and subverted in numerous instances, and most notably applied as a prototype for a visual programming language in *Visual Representation for Video Game Description Language*. The author invites any future Dagstuhl Seminar or reader to use it in their own work. Have fun and make it your own!

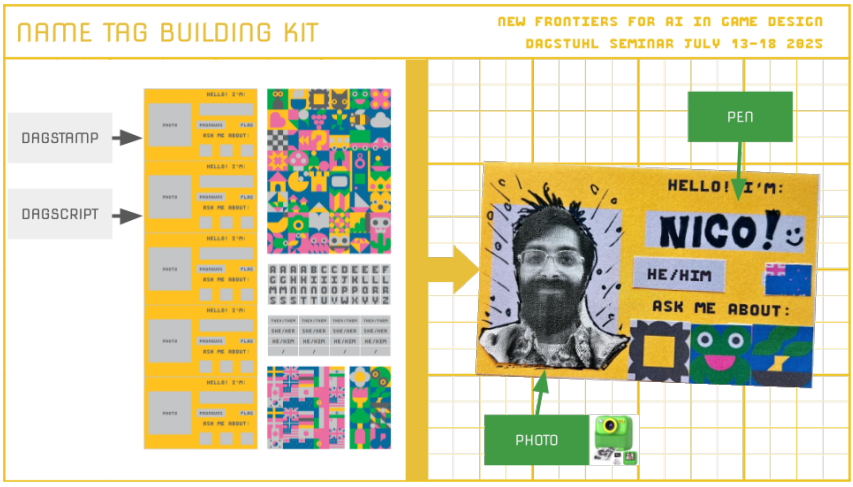


Figure 40 An example nametag used to teach attendees how to make their own.

Participants

- Claus Aranha
University of Tsukuba, JP
- Maren Awiszus
Viscom AG – Hannover, DE
- In-Chang Baek
Gwangju Institute of Science & Technology, KR
- June Bhartia
Télécom Paris, FR
- Rafael Bidarra
TU Delft, NL
- Brian Bucklew
Freehold Games – Walkerton, US
- Duygu Cakmak
Creative Assembly – Horsham, GB
- M Charity
University of Richmond, US
- Kate Compton
Vejle, DK
- Michael Cook
King's College London, GB
- Alena Denisova
University of York, GB
- Rémy Devaux
Punkcake Délicieux – Cenon, FR
- Alexander Dockhorn
University of Southern Denmark – Odense, DK
- Matthew J. Guzdial
University of Alberta – Edmonton, CA
- Emily Halina
University of Alberta – Edmonton, CA
- Max Kreminski
Midjourney – Santa Clara, US
- Antonios Liapis
University of Malta – Msida, MT
- Tiago Machado
IBM Research – Sao Paulo, BR
- Timothy Merino
NYU – New York, US
- Younès Rabii
Queen Mary University of London, GB
- Dipika Rajesh
University of California – Santa Cruz, US
- Emily Short
Oxford, GB
- Adam M. Smith
University of California – Santa Cruz, US
- Gillian Smith
Worcester Polytechnic Institute, US
- Florence Smith Nicholls
Queen Mary University of London, GB
- Anne Sullivan
York University – Toronto, CA
- Yuqian Sun
Royal College of Art – London, GB
- Nicolaas Vas
Billund, DK
- Sabine Wieluch
Universität Ulm, DE

Report from Dagstuhl Seminar 25301

Linguistics and Language Models: What Can They Learn from Each Other?

Anna Rogers*[1], Nathan Schneider*[2], Bonnie Webber*[3], A. Seza Doğruöz†[4], and Asad Sayeed†[5]

1 **IT University of Copenhagen, DK.** arog@itu.dk
2 **Georgetown University – Washington, DC, US.** nathan.schneider@georgetown.edu
3 **University of Edinburgh, GB.** bonnie.webber@ed.ac.uk
4 **Ghent University, BE.** as.dogruoz@ugent.be
5 **University of Gothenburg, SE.** asad.sayeed@gu.se

Abstract

An international group of 40 scholars in computational linguistics, natural language processing, and cognitive science was assembled to discuss the relationship between linguistics and contemporary language models. Over the course of the week, presentations and work sessions grappled with questions about how LMs can support linguistic research (either as a source of evidence, or as a tool); how linguistic knowledge can inform the design, interpretation, or application of LMs; and what framing is appropriate for the language functionality of LMs.

Seminar July 20–25, 2025 – https://www.dagstuhl.de/25301
2012 ACM Subject Classification Computing methodologies → Cognitive science; Computing methodologies → Natural language processing
Keywords and phrases cognitive modelling, language models, linguistic theory
Digital Object Identifier 10.4230/DagRep.15.7.187

1 Executive Summary

Nathan Schneider (Georgetown University – Washington, DC, US)
Anna Rogers (IT University of Copenhagen, DK)
Bonnie Webber (University of Edinburgh, GB)

Since the release of ChatGPT, language models (LMs) have stirred concerns in government, over the possibility that citizens will come to believe the textual and spoken output of such models. Similarly, they have caused panic in education, forcing a rethink of what students are learning and how to assess it. Of concern to us here, is whether LMs mean the end of computational and/or cognitive models of human language learning and language use. Does the practical success of LMs mean that computational linguistics (and perhaps even linguistics itself) is no longer relevant? Or are we missing problems with LMs that computational linguistics (and linguistics more generally) could help us both recognize and surmount?

* Editor / Organizer
† Editorial Assistant / Collector

Linguistics and Language Models: What Can They Learn from Each Other?, *Dagstuhl Reports*, Vol. 15, Issue 7, pp. 187–212
Editors: Anna Rogers, Nathan Schneider, Bonnie Webber, A. Seza Doğruöz, and Asad Sayeed
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

To have any hope of answering big questions about this technology, we need to foster interdisciplinary conversations and collaborations across the fields of machine learning, NLP, linguistics, and cognitive science. This Dagstuhl Seminar was organized to facilitate such conversations and collaborations among senior experts and rising stars. In particular, the five key questions were raised for discussion:

- What evidence, if any, do LMs provide about human language, world knowledge and/or cognition?
- How can LMs be used as tools for empirical research in linguistics?
- How can linguistics be brought to bear on interpreting the operation of LMs?
- How can linguistically-oriented perspectives enhance or complement LMs for greater reliability and robustness?
- What is the appropriate framing of LM-functionality, for scientists and the public?

An international group of 40 scholars in computational linguistics, natural language processing, and cognitive science was assembled for our week-long seminar. Commensurate with the broad questions raised in the seminar, participants were selected for their wide-ranging expertise on topics such as computational cognitive modeling and psycholinguistics; multilingual modeling and language variation; formal and functional aspects of language use; machine learning; LM interpretability; NLP for low-resource languages; applications and social impacts of language technologies; and philosophical underpinnings of modeling language.

The scientific program consisted of

- Eleven 20-minute talks raising perspectives and questions to inspire further discussion.
- Two rounds of working groups formed dynamically based on participant suggestions. The first set of groups held parallel meetings on Monday/Tuesday, each presenting a synopsis in a plenary session Tuesday evening. The second round of groups took place Wednesday morning and Thursday, reporting back in a Thursday evening plenary session.
- Friday morning was devoted to a plenary discussion of next steps, with about a dozen participants volunteering to organize follow-up initiatives to capitalize on some of the most fruitful conclusions of the working groups.

Abstracts from the talks as well as the working groups are reported below. In true Dagstuhl fashion, the formal scientific program was complemented by opportunities for socialization and recreation in and around the castle – the lively exchange of ideas and perspectives that began in the official sessions continued over meals, coffee breaks, nature hikes, and a sightseeing excursion to Trier.

Finally, a word of thanks from the organizers: We are grateful to all the attendees and the Dagstuhl staff who made the seminar an incredible experience. Special shoutouts go to Christina Schwarz for her administrative leadership; to participants A. Seza Doğruöz and Asad Sayeed, who agreed to serve as collectors for the final report; and to Asad also for his organizational assistance with the Wednesday social outing to Trier.

2 Table of Contents

3 Overview of Talks

3.1 Endangered Languages, Language Varieties, and LLMs

Antonios Anastasopoulos (George Mason University – Fairfax, US)

This talk explores the potential for interaction between documentary or descriptive linguistics on one side, and on NLP and LLM-focused researchers on the other. In this talk I'll argue the importance of documenting the wealth of linguistic diversity, and outline specific tasks amenable to technological intervention in the documentation process. Next, I will outline a proposal and show evidence from preliminary experiments for building language technologies in under-served languages, centered around careful small-scale data curation and on leveraging already-codified linguistic knowledge in the form of descriptive grammars.

3.2 Remarks on the Distributional Foundations of Language Models

Juan Luis Gastaldi (ETH Zürich, CH)

Linguistic distributionalism has been identified as a central theoretical principle explaining or justifying the success of neural LMs. However, there is a significant gap between the theoretical principles associated with distributionalism and the formal mechanisms governing current LMs. I proposed an interpretation of distributionalism based on current developments in category theory and type theory that could help bridge this gap and critically assess our knowledge about LMs. I introduced this topic with more general epistemological remarks concerning the kind of interpretability one can expect to achieve through this formal approach. Concretely, I defended the following claims:

1. LLMs have no a priori cognitive import
2. The empirical study of LLMs has no epistemological grounds
3. Distributionalism is the best theoretical candidate to study LLMs
4. Distributionalism is a corollary of structuralism
5. The general form of distributions is $\mathcal{M}\colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathcal{V}$
6. The general form of structures is $\mathcal{M}^*\colon \mathcal{V}^{\mathsf{C}^{\mathrm{op}}} \rightleftarrows (\mathcal{V}^{\mathsf{D}})^{\mathrm{op}}\colon \mathcal{M}_*$
7. This new structuralist formalism provides new representational tools for explainability and interpretability
8. Language models are culture models

3.3 Count me impressed

Adele Goldberg (Princeton University, US)

Constructions are learned pairings of form and function, at varying levels of complexity and abstraction. They are many, varied, context-dependent and interrelated to one another. The functions of constructions can involve subtle aspects of meaning, attitude, speech acts, and information structure. I will present two quite subtle and distinct aspects of language that are replicated by LLMs.

First, is an historical shift in word order amongst a cluster of conjunctions: English speakers used to say, uncles and aunts, nephews and nieces, and pa and ma, but the preference for these cases reversed to today's female-first order today, over the course of many decades (Goldberg & Lee, 2021). By training a GPT2LMHeadModel from scratch on 111GB of text, with closely related conjunctions filtering out, we demonstrate that the gradual shift is replicated by the model at 3 iterations.

Second, English speakers' judgments about the information structure of canonical sentences predicts independently collected acceptability ratings on corresponding "long distance dependency" (LDD) constructions, across a wide array of base constructions and multiple types of LDDs (Cuneo & Goldberg, 2023). To determine whether any LM captures this relationship, we probe GPT-4 on the same tasks used with humans and new extensions. Results reveal reliable metalinguistic skill on the information structure and acceptability tasks, replicating a striking interaction between the two, despite the 0-shot, explicit nature of the tasks, and little to no chance of contamination. A second study manipulates the information structure of base sentences and confirms a causal relationship: increasing the prominence of a constituent in a context sentence increases the subsequent acceptability ratings on an LDD construction (Cuneo et al., 2025). The findings suggest a remarkable relationship between natural and LM-generated English that make LMs a rich resource for testing theories of language.

3.4 A model for language models

Aurelie Herbelot (Denotation – Pritzwalk, DE)

In spite of their excellent performance on a range of benchmarks, LLMs (and other computational models) have been found to struggle with phenomena that require precise extensional knowledge. This talk highlights a range of quantification phenomena which, arguably, require some kind of representation of individual entities and events to be processed. I will suggest that quantification is a case where linguistic theory can be usefully integrated into computational systems and I will show what such integration can look like when expressing fundamental parts of model theory in vector spaces. I will also briefly demonstrate that the integration has benefits for standard lexical tasks when learning over small corpora. The second part of the talk asks to what extent this kind of representational integration can take place in LLMs, where there is much less control over the spaces that emerge in the course of training. I will present a real-world case study where a LM must be trained over limited data and could benefit from some form of limited exensional knowledge.

3.5 The Changing Roles of (Linguistic) Structure in Computational Linguistics

Mark Johnson (Macquarie University – Sydney, AU)

This talk describes the various roles that linguistic theory and structure have played in computational linguistics, and speculates about the role that they may play in the future. The closest relationship between linguistics and computational linguistics was probably with the Unification Grammars introduced in the 1980s, where the goal was to develop a computational model that implemented the linguistic theory. This close relationship proved impractical for scientific and sociological reasons that I'll describe, and since then the relationship has steadily weakened, as reflected by Jelinek's "when I fire a linguist . . . " quip and Sutton's "Bitter Lesson". I argue that the huge training data and long context windows of Deep Learning models makes it unnecessary to incorporate any specific linguistically-inspired parsing architecture into such models. While there are deep scientific questions about how LLMs "understand" human languages, their linguistic ability is sufficiently good for most practical tasks. Quite reasonably most current research focuses on the information content of the language LLMs generate, such as reducing hallucinations and improving instruction-following. Thus it seems the main opportunities for linguistics to contribute to modern computational linguistics are in model evaluation and explainability.

3.6 What kind of learning is in-context-learning? Evidence from psycholinguistics

Tom McCoy (Yale University, US) and Robert Frank (Yale University, US)

The field of psycholinguistics has developed fine-grained methods for using behavior to understand the mechanisms that underlie human language processing. In this talk, we will argue that such methods can also be used to analyze the language-processing mechanisms employed by large language models (LLMs). We will illustrate this point through a case study about in-context learning, the poorly-understood ability of LLMs to "learn" a task via examples presented to them in context, without explicit parameter updates. One line of research has claimed that in-context learning is functionally equivalent to gradient descent, a type of error-driven learning mechanism, but this claim remains controversial. As a new source of evidence in this debate, we draw a connection between in-context learning and structural priming, the psycholinguistic phenomenon in which people tend to produce sentence structures they have encountered recently. Structural priming literature has argued that human structural priming involves error-driven learning, a conclusion based on evidence from the inverse frequency effect (IFE) – a phenomenon in which an agent's behavior is influenced to a greater degree when presented with improbable examples as compared to more likely ones – which is a signature of error-driven learning. We show that the IFE is also present in LLMs performing in-context learning, which is evidence that in-context learning does indeed behave as an implicit type of error-driven learning. This work provides an example of how psycholinguistic methods can illuminate not only the aspects of grammar that LLMs have or have not captured (a direction that is common in existing work) but also the types of processing mechanisms that drive the language-processing abilities of LLMs.

3.7 Causal abstraction as a toolkit for developing linguistic theories

Christopher Potts (Stanford University, US)

Large language models (LLMs) are incredible new investigative tools for linguistic research; modern linguistics is based in distributional evidence, and LLMs are exceptionally powerful distributional learners. In this talk, I'll illustrate this potential using causal abstraction, which can characterize the abstract representations that LLMs have learned to use. Causal abstraction analyses reveal that LLMs have induced many of the core constructs posited by syntactic theories, while also suggesting new factors that may be relevant to such theories. I'll close with a discussion of the poverty of the stimulus. LLMs reveal that the distributional evidence is richer than previously assumed, which should lead us to reassess arguments that specific phenomena cannot be learned from the available data.

3.8 Linguistics from First Principles and LLMs

Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA)

LLMs present an opportunity for linguistics. Instead of verifying whether linguist-annotated representations are present in these models, we can reverse the question: If we were to rely on foundational principles of linguistics, can we extract linguistic representations from the internals of LLMs? In this talk, we will use foundational principles for syntax and semantics to directly extract dependency structures and model-theory-based semantics from LLMs.

In the latter part of the talk, we will focus on how to use LLMs to gain insights into their linguistic representations, directly using prompts. We will contrast different settings: probabilities versus meta-linguistics, and instruction-following versus base models. Finally, using recent reasoning models like DeepSeek-R1, which are trained to discover chains of thought that could lead to the correct answer, we will analyze their reasoning processes on linguistic stimuli.

3.9 What are Large Language Models Models Of?

Philip Resnik (University of Maryland – College Park, US)

Should we be thinking of LLMs as cognitive models for human language processing? This talk went beyond the "argument from amazingness" -- that LLMs are so good at human language that they must be doing something similar to what we do -- to a more careful and critical assessment of what it means to model human language processing, and why it might or might not make sense to view LLMs as models in that sense.

The discussion is organized in terms of Marr's levels of explanation. At the implementation level, we argue that, to whatever extent neo-connectionist models like transformers remain "biologically inspired", that inspiration comes from a version of neurobiology that is at least

a half-century out of date. As such, LLMs are a poor candidate for implementation-level models of human language processing, especially in contrast to the current and actively growing new generation of implementation-level models of processing that actually are in line with current neuroscientific knowledge.

At the algorithmic/representation level, it's worth acknowledging that aspects of deep learning are genuinely a game-changer with respect to theories of human language processing, notably in the success of representation learning as an alternative to manually constructed representations. However, the feed-forward architecture of LLMs is misaligned with the increasing body of evidence that relevant aspects of human perception and cognition are predictive in nature – cf. the Bayesian Brain hypothesis and its realization in predictive coding models. Moreover, there is no basis for the widely held belief that two systems solving the same problems must necessarily have similar solution mechanisms: nature provides numerous counter-examples where different organisms solve the same problems in vastly different ways. And if you look at how model organisms are chosen in medical research, a good model is not one that just manifests a similar behavior or phenomenon, there are also independent reasons for believing what we learn from a model organism will transfer over to humans. For example, mice and rats are used in cancer and cardiovascular disease research respectively, not the reverse, because we know mice have similar oncogene pathways to humans and rats have similar cardiovascular physiology.

At Marr's computational theory level, even LLMs' amazing capabilities in language input/output behavior don't constitute a convincing argument for their treatment as cognitive models. Two systems with identical input/output can still differ drastically in the way they work – e.g., exactly the same input and output pairs are generated by iterative and recursive implementations of the factorial function. We care about computation-level theories in large part because they are a step toward underlying mechanistic understanding. Even if LLMs could achieve the right input/output behavior across a wide range of human language phenomena, there's no reason to believe that there will then be a way to get from the one-size-fits-all, homogenous computation-level framework to anything resembling the evolved solution inside human heads. In that regard, faith in the cognitive relevance of LLMs is ironically very similar to the Chomskian approach in theoretical linguistics, both fixated on achieving an elegant and uniform solution to problems that nature solves messily – nature being a scavenger building on existing capabilities, not a designer refactoring a system to maintain its elegance and uniformity.

Ultimately, then, we find that LLMs don't offer a convincing case as cognitive models, at any level of explanation. What they do offer – in a truly game changing way – is a new kind of support tool for developing actual cognitive models: an ability to provide missing jigsaw puzzle pieces in models that require distributed representations, proxies for world knowledge, stand-ins for plausible inference, and much more. One can lean into the amazing capabilities of LLMs for developing cognitive models without having to buy into the idea that they constitute models in and of themselves.

3.10 Combining Large Language Models and Symbolic Systems for Logical Inference

Mark Steedman (University of Edinburgh, GB)

The traditional view of Natural Language Processing (NLP) distinguishes two components: The Grammar, which is recursive and logic-related, and supports syntactic analysis and semantic interpretation; and The Model, which is probabilistic, and assigns measures of similarity of association or context to symbols and sequences.

This distinction is parallel to a widespread view of psychologists that human understanding proceeds via "Dual Processes", involving both systems of "Type 2", consisting of high-precision but slow symbolic systems, and "Type 1" systems, consisting of fast, sloppy, imprecise but high-recall systems such as Finite-State Transducers (FST) or Language Models (LM). Type 1 systems are usually "good enough", but can be monitored or augmented by Type 2 systems via channels of very limited bandwidth.

The primary purpose of the Grammar/Type 2 system of NLP is to support sound logical inference, while the purpose of the Model/Type 1 system is to limit ambiguity in mapping strings to meanings and vice versa, via the context of utterance.

Large Language Models (LLM) consisting of up to a trillion parameters and trained on trillions of words of human generated text have recently proved extremely useful in a number of NLP tasks such as question answering and summarization. However, such tasks involve inference from the form of statements in documents to possible statements in the output. The talk will review capabilities of LLMs for Natural Language Inference (NLI) in such applications.

We can think of LLMs as a hypersphere of vector embeddings with only a few hundred dimensions of associative similarity, algorithmically compressed by dimension reduction from the vaster dimensionality and empty sparsity of the raw association space obtained from text.

Crucially, there is known to be a gradient of generality on each axis from very general terms at the center of this hypersphere to more specific terms at the periphery. It will be important to what follows to recall that terms that stand in the relation of a generalization ore known to also form a partial ordering of frequency in text.

The talk will argue that, once certain biases in the human-constructed NLI datasets and in the LLM themselves are controlled for, LLMs perform quite badly on pure NLI tasks, despite fair recall, showing poor precision or false-positive conclusions ("hallucinations"). In particular, LLM are prone to two biases that are inherent in the text training data, namely an Attestation Bias stemming from the fact that those data concern facts, and a Frequency Bias, stemming from the fact that some things are more written about than others.

The paper will argue instead for the unsupervised extraction of "Entailment Graphs" (EG) or probabilistic NLI networks via detection using machine reading of probabilistic typed entailment relations such as that buying events imply ownership states events. EG are higher-precision, and our methods scale. However, Zipf's Law means that they are subject to an intrinsic curse of sparsity.

The paper presents results from the Edinburgh group developing hybrid systems that combine the high precision of EG with the high recall of LLM by exploiting the biases themselves of the latter. We show that the Frequency gradients in the latter can be leveraged to "smooth" entailment premises that are missing from EG via nearest LLM neighbors that

are present in the EG, with the entailment graph doing the rest of the work. We also exploit the Attestation Bias of LLM to find entailments de novo by eliciting attested analogs of the Premise, then querying the attestation of parallel Hypotheses.

Our conclusion is that the future of NLI and related applications lies with such cautious hybridized combination of the Precision of symbolic Type 2 inference systems with the Recall of LLM Type 1 systems.

3.11 Can (and should) an AI do science?

Adina Williams (Meta Platforms – New York, US)

LLMs are now good enough to generate text that can plausibly pass as a scientific paper. Papers have been used as evidence of scientific advancement, both for hiring/recruiting and for progressing in our field's scientific goals. The incentives of our field make it likely that people will use them for generating papers. Since an LLM is not a part of our community and we cannot trust that the texts it outputs is good evidence of science having been done, LLMs may contribute to undermining our scientific community especially our peer review system which we all rely on to scale scientific trust beyond tight friend and or collaborator networks.

4 Working groups

4.1 Multilingualism and low-resource languages

A. Seza Doğruöz (Ghent University, BE), Verena Blaschke (Ludwig-Maximilians-Universität München, DE), David Adelani (MILA – Montreal, CA), Lori Levin (Carnegie Mellon University – Pittsburgh, US), Xixian Liao (Barcelona Supercomputing Center, ES), Joakim Nivre (Uppsala University, SE), Alexis M. Palmer (University of Colorado – Boulder, US), Gözde Gül Şahin (Koç University – Istanbul, TR), and Amir Zeldes (Georgetown University – Washington, DC, US)

Although there are ∼7,000 languages in the world, most research in computational linguistics and natural language processing (NLP) focuses on English [11, 5, 19]. To study multi- and cross-lingual aspects of language models, multilingual datasets and/or datasets in low-resource languages are key. Nevertheless, many datasets cover English and multilingual datasets often contain content specific to some cultures more than others. Many multilingual datasets are based on translations, requiring caution with respect to translationese [17]. Some datasets might also disregard linguistic variation, especially within non-standardized language varieties. A lack of standardization of formats, tools, and APIs makes it harder to compare datasets (or to analyze models across datasets). Additionally, quality control and reachability of a responsible human maintainer are important for the use of datasets by others.

It is also difficult to evaluate multilingual models meaningfully. Massively multilingual benchmarks can be biased (as mentioned above). Intrinsic, task-independent evaluation that is comparable across languages remains a challenge, and intrinsic evaluation results are not necessarily correlated with real-world performance. Automatic metrics can be biased against certain language types and they can be brittle towards variation [20, 1]. If they are learned from data they may simply not cover low-resource languages. Possible steps forward include taking inspiration from the Multidimensional Quality Metrics framework [13], categorizing evaluation metrics for different languages, monitoring existing evaluation sets for content and language issues, and prioritizing speaker community involvement in the creation of new datasets.

Insights from linguistic typology could also be used for interpreting multilingual language models. Methods from interpretability research (e.g., probing, training sparse auto-encoders, or designing causal interventions; see [3], and [16]) could be used to investigate whether the internal representations of multilingual models organize linguistic information in a way that it mirrors typological patterns or shows systematic differences between language-specific vs. cross-lingual representations (e.g., [6], and [21]).

Although many assumptions in NLP research may not necessarily hold for low-resource languages, and standard parameter and/or tokenization settings might be suboptimal for many of them, these challenges may still be disregarded for processing and studying these languages. Language models that are of sufficient quality for some downstream applications might not meet the quality standards needed for archival-quality documentation that can support linguistics research and language revitalization [10]. Language use can also be different than frequently assumed. For example, some speaker communities are multilingual and utilize more than more language/dialect on a daily basis [9]. Furthermore, textual data is also prioritized over informal, spoken data [7] although there are many languages that are only spoken (not written) and/or they may not have standardized orthographies.

Lastly, it is important to support and involve speaker communities and field linguists when creating NLP tools for low-resource languages [8]. Community members could be consulted before building NLP products for their languages to ensure that their needs and preferences are met [12, 14, 4]. Furthermore, there are many opportunities (especially in low resource languages) for building NLP tools that can help field linguists (e.g., for transcribing speech data, or OCR tools for extracting text from images) if they are built with their needs in mind [15, 10]. Additionally, NLP tools could be used to aid with analyzing language data. A task like inferring linguistic structures from data (meta-linguistic reasoning [18]) might be of interest to both linguists and computer scientists.

References

1 Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore, December 2023. Association for Computational Linguistics.

2 Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, November 2020. Association for Computational Linguistics.

3 Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

4 Steven Bird and Dean Yibarbuk. Centering the speech community. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

5 Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics.

6 Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

7 Grzegorz Chrupała. Putting natural in natural language processing. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827, Toronto, Canada, July 2023. Association for Computational Linguistics.

8 A. Seza Doğruöz and Sunayana Sitaram. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In Maite Melero, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France, June 2022. European Language Resources Association.

9 A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. A survey of code-switching: Linguistic and social perspectives for language technologies. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online, August 2021. Association for Computational Linguistics.

10 Luke Gessler, Alexis Palmer, and Katharina Von Der Wense. Understanding the gap: an analysis of research collaborations in NLP and language documentation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 867–877, Vienna, Austria, July 2025. Association for Computational Linguistics.

11 Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.

12 Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. Not always about you: Prioritizing community needs when developing endangered language technology. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland, May 2022. Association for Computational Linguistics.

13 Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib.

14 Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada, July 2023. Association for Computational Linguistics.

15 Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaum. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12, 2018.

16 Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, pages 1–48, 09 2025.

17 Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. Culture is everywhere: A call for intentionally cultural evaluation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19156–19168, Suzhou, China, November 2025. Association for Computational Linguistics.

18 Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. PuzzLing Machines: A Challenge on Learning From Small Data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online, July 2020. Association for Computational Linguistics.

19 Anders Søgaard. Should we ban English NLP for a year? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

20 Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. Dialect-robust evaluation of generated text. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada, July 2023. Association for Computational Linguistics.

21 Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

4.2 Accommodation and Social Aspects of LLMs

A. Seza Doğruöz (Ghent University, BE), David Adelani (MILA – Montreal, CA), Antonios Anastasopoulos (George Mason University – Fairfax, US), Verena Blaschke (Ludwig-Maximilians-Universität München, DE), Aurelie Herbelot (Denotation – Pritzwalk, DE), Yu-Yin Hsu (Hong Kong Polytechnic Univ., CN), Xixian Liao (Barcelona Supercomputing Center, ES), Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA), Philip Resnik (University of Maryland – College Park, US), Anna Rogers (IT University of Copenhagen, DK), Gözde Gül Şahin (Koç University – Istanbul, TR), Asad Sayeed (University of Gothenburg, SE), Noah A. Smith (University of Washington – Seattle, US), and Bonnie Webber (University of Edinburgh, GB)

LLMs are widely used across domains but there are relatively less insights into the variation and change within languages in terms of context and speakers/users involved in communication. Socio-demographic characteristics of the speakers and context in human-human communication influence the language use ([4]). Humans share common ground and accommodate each other to achieve common communication goals ([5, 2]). Since there is not always a common ground between LLMs and a human ([3]), it is difficult to measure and evaluate accommodation.

However, not all humans approve of LLM's accommodation in language use (e.g., attitudes of dialect speakers toward LLMs by [1, 6]). It is also the case that for certain domains (e.g., mental health), it is favorable if the LLMs do not always accommodate the users.

Possible applications of this research direction include personalization of LLMs for specific users. However, there is a need for more research to understand the goals and social aspects of human-human interaction across contexts and among users with varying socio-demographic characteristics.

References

1 Blaschke, V., Purschke, C., Schuetze, H., Plank,B. (2024). What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 823-841, Bangkok, Thailand. Association for Computational Linguistics.

2 Herbert H Clark. 1996. Using language. Cambridge university press.

3 Doğruöz, A.S. & Skantze, G. (2021). How "open" are the conversations with open-domain chatbots? A proposal for Speech Event based evaluation. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 392–402, Singapore and Online. Association for Computational Linguistics.

4 Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F. (2016). Computational Sociolinguistics: A Survey. Computational Linguistics, 42(3):537–593.

5 Pickering MJ, Garrod S. (2021) Understanding Dialogue: Language Use and Social Interaction. Cambridge University Press.

6 Sandoval, S.C, Acquaye, C., Cobbina, K.A., Teli, M.N., and Daumé, H. (2025). My LLM might Mimic AAE – But When Should It?. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5277–5302, Albuquerque, New Mexico. Association for Computational Linguistics.

4.3 Goals of Linguistic Theory

Robert Frank (Yale University, US), Katherine Demuth (Macquarie University - Sydney, AU), Juan Luis Gastaldi (ETH Zürich, CH), Mark Johnson (Macquarie University - Sydney, AU), Najoung Kim (Boston University, US), Roger Levy (MIT - Cambridge, US), Siva Reddy (MILA - Montreal, CA & McGill University - Montreal, CA), Philip Resnik (University of Maryland - College Park, US), Anna Rogers (IT University of Copenhagen, DK), Rachel Rudinger (University of Maryland - College Park, US), Nathan Schneider (Georgetown University - Washington, DC, US), Mark Steedman (University of Edinburgh, GB), Tiago Torrent (Federal University of Juiz de Fora, BR), and Adina Williams (Meta Platforms - New York, US)

This group focused on questions at the foundation of the science of language. Our first focus for discussion sought to identify the goals of linguistic theory and its objects of study. There was broad agreement that the focus of the field should be on "explaining patterns of language," but there was broad ranging discussion about what constitutes such a pattern. On the one hand, participants agreed that patterns of sound (phonology), word structure (morphology) and sentence structure (syntax) are central, as are the ways in which meaning is encoded in the words of language (lexical semantics) and the ways such meanings can be put together to produce interpretations of larger units (compositional and discourse semantics). Patterns of language use (pragmatics) and the fit of linguistic form to social and political context were also mentioned as important areas of study, but there was some belief that social and political factors governing language use may lie outside the core of language behavior and would be better understood through the interaction of the language faculty with other cognitive capacities.

This latter point led to a discussion about the ways in which the field has partitioned linguistic and non-linguistic cognition. We reviewed Fodor's ([4]) notion of modularity, according to which language knowledge, as a cognitive module, should be be cleanly separated from non-modular systems of thought involved in the use of world knowledge. LLMs don't appear to make such a division between linguistic and non-linguistic cognition, and this led to a discussion about whether this might be taken as a reason to reconsider the modularity assumption. One empirical reason to move in this direction might come from the line of work in the psycholinguistics in the 1980s and 90s (e.g., [7]) which showed that it was difficult to distinguish the processing impact of linguistic and non-linguistic information in sentence processing (though there are ways of reconciling these results with modularity – cf., [1]). Recent work on LLMs as models of sentence processing (e.g., [8]) seems to point to the conclusion that LLMs transcend human performance, in that they do not show sensitivity to the same limitations that humans do. Consequently, the lack of modularity in LLMs might not be reason to doubt its presence in human processing.

Our discussion then moved to the means by which linguistic theories should be judged. General scientific desiderata were discussed, including falsifiability and testability of predictions and insight and explanatory value. Other factors that came up were more specific to linguistics: the explanation of the capacity for systematic generalization (e.g., to novel forms) and the ability to account for typological patterns (including entailments from one type of grammatical property to another and the non-existence of certain types of grammars). Finally, there was broad agreement that theories of language should engage with patterns of behavior in comprehension and production, as well as with the characterization of developmental trajectories during language acquisition.

With these foundations in mind, we returned to the question of assessing the viability of LLMs as theories of human language. Group members agreed that LLMs are clearly impressive in their ability to carry out certain tasks. However, there was broad agreement that it is difficult indeed to know how to evaluate their success scientifically. A number of open questions were raised about what could in principle be learned from LLMs: what is the nature of their inductive bias that contributes to their success in language learning? what can LLMs tell us about the relevant data structures for language? Much of the work on LLM interpretability in the linguistic vein has focused on finding analogs of known linguistic structures in the learned representations of LLMs (e.g., [6]). An especially interesting example of this arose in the talk at the workshop by Chris Potts ([2]) pointing to the convergence in representation across the wh-movement constructions (C[3]). If this is the kind of finding we repeatedly find, it points to the validity of the abstractions of current linguistic theory, but also says that the model structure, training data and gradient descent-driven learning process are sufficient to yield relevant abstractions without innate guidance.

Our final day of discussion focused on the problem of evaluation metrics for LLMs as models of language structure. The problem stems from the fact that traditional ways of thinking about grammar learning assume discrete languages, yet LLMs are probabilistic models, which do not simply rule strings in and out of a language, bur rather provide a distribution over such strings. To judge the fit between and LLM and a linguistic pattern, a variety of approaches have been explored. Most simply, one can compare the probability assigned to sentences in a minimal pair, with the expectation that the grammatical sentence should receive higher probability than the ungrammatical one. A number of problems were discussed for this approach: how do we determine what constitutes a minimal pair? How do we deal with other facts that can impact assigned probability (e.g., frequency effects, semantic plausibility, etc.), but which are plausibly orthogonal to language structure itself? More complex approaches attempt to provide a fixed score for sentence, by normalizing the assigned probability for sentence length and word frequency ([5]). We began to explore a third possibility based on the idea that learning requires demonstrating use of relevant latent variables or structures. We discussed a number of ways that such latent structures could be identified. One direct way would be via interpretability methods, but the participants expressed varying degrees of skepticism as to whether we can have confidence in current methods. An alternative approach began to emerge by examining the impact the relevant abstractions would have on the distributions generated by a model containing them.

References

1 Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. Cognition, 30(3), 191–238.

2 Boguraev, S., C. Potts & K. Mahowald. (2025). Causal interventions reveal shared structure across English filler-gap constructions. Proceedings of EMNLP.

3 Chomsky, N. (1977). On Wh-Movement. In P. Cullicover, T. Wasow, & A. Akmajian (Eds.), Formal Syntax (pp. 71-132). New York Academic Press.

4 Fodor, J. A. (1983). Modularity of Mind: An Essay on Faculty Psychology. Cambridge, Massachusetts: MIT Press.

5 Lau, J.H., Clark, A. and Lappin, S. (2017), Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. Cogn Sci, 41: 1202-1241.

6 Manning, C.D., K. Clark, J. Hewitt, U. Khandelwal, & O. Levy. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision, Proc. Natl. Acad. Sci. U.S.A. 117 (48) 30046-30054.

7 Tanenhaus, M. K., Dell, G. S., & Carlson, G. (1987). Context effects in lexical processing: A connectionist approach to modularity. In J. L. Garfield (Ed.), Modularity in knowledge representation and natural-language understanding (pp. 83–108). The MIT Press.
8 Van Schijndel, M. & T. Linzen (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. Cognitive Science, 45(6), e12988.

4.4 Tools for exploring linguistic theories

Lori Levin (Carnegie Mellon University – Pittsburgh, US), Adele Goldberg (Princeton University, US), Yu-Yin Hsu (Hong Kong Polytechnic Univ., CN), Najoung Kim (Boston University, US), Tom McCoy (Yale University, US), Joakim Nivre (Uppsala University, SE), Alexis M. Palmer (University of Colorado – Boulder, US), Jakob Prange (Universität Augsburg, DE), Rachel Rudinger (University of Maryland – College Park, US), Nathan Schneider (Georgetown University – Washington, DC, US), and Tiago Torrent (Federal University of Juiz de Fora, BR)

The research question

This group investigated what kind of tools could allow a linguist to visualize the internal representations of Large Language Models (LLMs) so that the linguists could see whether significant linguistic phenomena show up in the internal representations of LLMs and study their representation in LLMs.

(Chris Potts and Mark Johnson mentioned some such tools in their plenary talks.)

The linguistic phenomenon

The group considered three linguistic phenomena:

1. The distinction between arguments and adjuncts
2. The distinction between new and presupposed components of a sentence
3. The distinction between metaphorical and non-metaphorical components of a sentence

The group decided to focus on the distinction between arguments and adjuncts. In the sentence *Alex ate the cake hungrily in Dagstuhl at 3:30pm*, *Alex* and *cake* are arguments of the verb *eat*, whereas *Dagstuhl* and *3:30pm* are adjuncts.

The argument-adjunct distinction plays a central role in many different linguistic theories. In Cognitive Grammar and Frame Semantics arguments may be conceived of as core participants in a scene, which can be brought to mind by either the valency properties of a predicator or by a (non-lexical) construction. In X-bar Theory arguments are closer to the head in a non-recursive projection; adjuncts are farther from the head in a recursive projection. In Case and Theta theory a head can assign case and semantic role to an argument. In Tree Adjoining Grammar arguments are attached to projections of heads via a substitution operation in contrast to adjuncts, which are attached by an adjunction operation. In Categorial Grammar heads are functions that apply to arguments, while adjuncts are functions that apply to the phrases they attach to. We would like to know whether arguments and adjuncts feature as significantly in the internal representations of LLMs as they do in linguistic theory.

A plan

Step 1: Create a dataset of behavioral diagnostics. There are many behaviors that distinguish arguments from adjuncts, including the following:
Omissability (adjuncts are more omissible)
I gave books to students on Thursday.
I gave books to students. (better)
I gave books on Thursday. (worse)
Repeatability (adjuncts are repeatable)
I gave books to people to students on Thursday. (worse)
I gave books to students on Thursday at 3:00. (better)
Selectional Restrictions (predicates impose selectional restrictions on their arguments)
I ate cake (normal)
I ate ideas (metaphorical)
Ideas eat cake (metaphorical)
Order in noun phrases (arguments before adjuncts)
A student of linguistics with long hair (Radford textbook) (better)
A student with long hair of linguistics (worse)
Order in sentences arguments before adjuncts
I gave books to students on Thursday. (better)
I gave books on Thursday to students. (worse)

Step 2: Test whether the LLM replicates human behavior, for clear-cut and non-clear-cut cases. Replicating human behavior might be reflected in perplexity, surprisal, or some other metric in an LLM. We would need to see whether the sentences that are judged as better and worse by humans in the examples above are also measurable as better and worse by some metrics in an LLM.

The argument-adjunct distinction, in spite of its central role in linguistic theory, is notoriously fuzzy around the edges. There are many diagnostics that give ambiguous results. For example, optionality is a defining characteristic of adjuncts, but some adjunct-like things are required in some constructions and many arguments are optional. In addition, humans cannot always make clear judgements about the acceptability of sentences. It is important to know whether LLMs behave like humans in these non-clear cut cases.

What if LLMs do not replicate human behavior? We will experiment with modifications of LLMs to see what kinds of adjustments affect their behavior with respect to the argument-adjunct distinction.

Step 3: Interpretability. We will examine components of LLMs such as circuits and subspaces of hidden states to see which components are implicated in their behavior toward arguments and adjuncts. The major research questions will be where to look, how to avoid confounds, and whether we will need to develop new interpretation techniques if we see nothing at first.

Step 4: See what those internal components do in non-clear-cut cases.

Step 5: Bring the lessons we learn from LLMs' representation of arguments and adjuncts back to linguistic theory.

Inspiring new human experiments: If our methodology for studying interpretability of the argument-adjunct distinction is unsupervised, it could lead us to discover new features or mechanisms that linguists were previously unaware of. We could then develop experiments that would verify whether similar things can be found in humans

Gaining new insights into characterizing the argument/adjunct distinction: If LLMs capture human behavior we might conclude that what they are doing provides one possible

account of how the behavior can be captured. That is, we might use LLMs to discover mechanisms for representing the argument-adjunct distinction that were not found by traditional methods such as corpus studies and human psycholinguistic experiments.

We would especially want to know whether the non-clear-cut cases exhibit a blend of the argument and adjunct signatures, or combinations of discrete sub-pieces of these, or something totally different? If it is a blend, how should this be reflected in updated linguistic theories?

4.5 The (Im)possible Languages Group

Christopher Potts (Stanford University, US), Marie-Catherine de Marneffe (UC Louvain-la-Neuve, BE), Katherine Demuth (Macquarie University – Sydney, AU), Robert Frank (Yale University, US), Juan Luis Gastaldi (ETH Zürich, CH), Hagen Blix, Coleman Haley (University of Edinburgh, GB), Mark Johnson (Macquarie University – Sydney, AU), Roger Levy (MIT – Cambridge, US), Kyle Mahowald (University of Texas – Austin, US), Mark Steedman (University of Edinburgh, GB), and Adina Williams (Meta Platforms – New York, US)

The (Im)possible Languages Group sought to understand the nature of the distinction between possible and impossible human languages, and to explore ways in which Large Language Models (LLMs) might help us explore the issue. Like many past researchers in this area, we were partly inspired by Jorge Luis Borges' short story "The library of Babel" (see [3]), but we infused this with (perhaps less erudite) references to the Mission Impossible movie franchise ([2]) and the Cole Porter musical Anything Goes ([6]).

We centered our discussion around the following claim: There are limits on the set of languages humans can acquire and use as full-fledged natural languages.

What are the nature of these limits? We identified five classes: (1) formal learnability and complexity, (2) expressivity, (3) efficiency, (4) historical/social, and (5) cognitive. We ventured that everyone accepts that (1)-(4) contain robust constraints. The central question is whether there is anything left for the cognitive class (5) that has, often implicitly, been the focus of debates in the field. (See also [4], chapter 3.)

We also agreed that talking about a "set of languages" in this context is misleading. Most or all of the limiting factors (1)-(5) will be a matter of degree, and acquisition itself might be partial and fluid. Thus, the relevant notions are actually quite fuzzy. Nonetheless, there will be clear enough cases (even if there is a large gray area), and so we feel we can talk informally about sets of languages.

What does the phrase "acquire and use as full-fledged natural languages" mean? Here, we did not achieve complete consensus. We agreed that, for a language to count as possible, it must be learnable and usable by ordinary humans without auxiliary tools, and it must be expressive enough to serve as a medium of human thought and communication. Other potential criteria include whether a language could have a lasting and stable community of fluent speakers, or whether it is processed in the brain with the same neural correlates we associate with other natural languages.

The criterion of neural correlates led us to pose a thought experiment. Identical twins Avery and Blake are separated at birth. Avery learns a regular old natural language, while Blake grows up in a community that speaks a typologically unusual system with counting-based rules that linguists would consider "impossible". Now suppose we scan their brains during language use and find that Avery's processing patterns are the expected ones for language whereas Blake's are ones we associate with different cognitive processes. Do you conclude that Blake's language isn't a natural language, or do you conclude that the neuroscience was wrong? Members of our group gave different answers.

We then turned our attention to the role that LLMs might play in these investigations. Using the above framework, we determined that the Shuffle language experiments of [2] are focused on contraints that are not in the cognitive group (5) above, but rather can be attributed to issues of formal complexity or expressivity. However, their Hop languages seem clearly to be oriented around a potential cognitive constraint.

[1] is an insightful critical review of [2]. We focused on Hunter's concern that Kallini et al.'s counting-based rules should be compared against something of similar complexity that crucially makes reference to constituency. To this end, Hunter sketches a "Sister Hop" language in which an agreement marker is placed after the constituent that is the sister of the verb. This is a realistic pattern that, he argues, is a fairer comparison to the Word Hop rule of Kallini et al., which places the agreement marker 4 words after the verb regardless of consituency. The group agreed that Hunter's language is hard to operationalize because of a lack consensus on constituency and/or a lack of parsers that could perfectly implement the correct theory at scale. However, we hit upon an alternative: use a language where Sister Hop is the actual pattern, and then derive the unnatural language from pattern. English verb–particle cases provide one such pattern; in phrases like "Pick the book on the table up", the particle appears immediately to the right of the sister of the verb. The (by hypothesis) impossible variant would place the particle N words from the associated verb, with N likely set to 4. This seems like a very promising basis for an informative experiment.

To close, we asked how we explain the result, now reproduced in a few papers ([5], [6]; but see [7]), that LLMs do distinguish at least some possible languages from impossible variants. One idea is that the AI community has evolved towards architectures that favor possible languages. This evolution could even extend to hyperparameter choices people have inherited over time from prior work and now use without much reflection. The precise factors remain to be discovered. Another idea is that complexity notions might organize the languages considered thus far experimentally. This would mean that we haven't yet evaluated any of the "cognitive" constraints that are the most meaningful to the debate.

References

1 Hunter, Tim. 2025. Kallini et al. (2024) do not compare impossible languages with constituency-based ones. https://direct.mit.edu/coli/article/doi/10.1162/coli_a_00554/128121

2 Kallini, Julie; Isabel Papadimitriou; Richard Futrell; Kyle Mahowald; and Christopher Potts. 2024. Mission: Impossible Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 14691-14714. Bangkok, Thailand: Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.787/

3 Moro, Andrea. 2015. The Boundaries of Babel: The Brain and the Enigma of Impossible Languages, 2d edition. MIT Press.

4 Nefdt, Ryan M. 2024. The Philosophy of Theoretical Linguistics. Cambridge University Press.

5 Xu, Tianyang; Tatsuki Kuribayashi; Yohei Oseki; Ryan Cotterell; and Alex Warstadt. 2025. Can Language Models Learn Typologically Implausible Languages? https://www.arxiv.org/abs/2502.12317
6 Yang, Xiulin; Tatsuya Aoyama; Yuekun Yao; and Ethan Wilcox. 2025. Anything Goes? A Crosslinguistic Study of (Im)possible Language Learning in LMs. https://arxiv.org/abs/2502.18795
7 Ziv, Imry Ziv; Nur Lan; Emmanuel Chemla; and Roni Katzir. 2025. Large Language Models as Proxies for Theories of Human Linguistic Cognition. https://arxiv.org/abs/2502.07687

4.6 "Post-cephalopod" natural language understanding

Asad Sayeed (University of Gothenburg, SE), Ryan Cotterell (ETH Zürich, CH), Marie-Catherine de Marneffe (UC Louvain-la-Neuve, BE), Aurelie Herbelot (Denotation – Pritzwalk, DE), Najoung Kim (Boston University, US), Christopher Potts (Stanford University, US), Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA), Rachel Rudinger (University of Maryland – College Park, US), Bonnie Webber (University of Edinburgh, GB), and Ethan Wilcox (Georgetown University – Washington, DC, US)

After the publication of Emily Bender and Alexander Koller's ACL award-winning 2020 article "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data", we decided it was time to discuss whether the debate it instigated is still current in the face of recent developments in LLM technology, or whether there are cogent objections to their layout of debate. This is relevant, because the "Octopus Paper" still to a large extent sets the tone for claims of machine understanding in the face of chatbots that produce output that is usually understandable in conversational context and that humans widely consider plausible that a human could have written. That is, it rejects the idea that meaning can be acquired simply through the manipulation of form (textual representations) alone, which is a key characteristic of the text-only chatbots that were prevalent when the paper was published.

Discussion proceeded through an examination of a common claim: that the Octopus Paper was written in a manner that resembled classical arguments against the possibility of considering machines to be human-like intelligences, most particularly the famous Searle "Chinese Room" Gedankenexperiment. The group concluded that the Octopus Paper makes assumptions about its core allegorical scenario, that of an octopus clandestinely intervening in the conversations of two island-dwellers communicating by wire, that are actually incompatible with the Searlean scenario. Bender and Koller, most notably, do not believe that symbol streams are sufficient for behavioural perfection (unlike Searle), but they do allow that symbol manipulation could actually represent understanding under the right conditions (also unlike Searle).

The group noted that Bender and Koller would allow that the machine would have "understood" if there were evidence of grounding, and that that grounding could be as thin as multimodal content other than text (e.g., images). What has happened since the Octopus Paper is a vast expansion of multimodal models that do just that. However, they often accomplish this also using a Transformer architecture that also depends crucially on conversion of multimodal content into abstract symbols. If symbol-manipulation (in the sense of text) is not in itself meaning, then how is "multimodal" symbol-manipulation a more authentic way to characterize meaning?

The group discussed a potential resolution to this problem that consisted of the following:

- Taking an internalistic view of meaning representation: understanding as the correspondence between language and internal representations.
- Accepting that "understanding" is a continuum, with some entities understanding little, and some understanding much more.

In this way, we can say that a simple household cleaning robot could "understand" in some sense, but it is a very simple form of understanding.

If this concept of understanding is unsatisfactory to some (due to questions of qualia, grounding, embodiment, and so on), the problem is that there is essentially an infinite regression of potential objections, with a new objection surfacing every time an old one is satisfied. E.g.: How many aspects of human experience need to be represented? Is simulating biology enough? And so on.

The group resolved to write an article assessing how to incorporate the current state of technology into the on-going debate.

4.7 LLMs and memory

Amir Zeldes (Georgetown University – Washington, DC, US), Ryan Cotterell (ETH Zürich, CH), Yu-Yin Hsu (Hong Kong Polytechnic Univ., CN), Mark Johnson (Macquarie University – Sydney, AU), Siva Reddy (MILA – Montreal, CA & McGill University – Montreal, CA), Philip Resnik (University of Maryland – College Park, US), Anna Rogers (IT University of Copenhagen, DK), Gözde Gül Şahin (Koç University – Istanbul, TR), and Ethan Wilcox (Georgetown University – Washington, DC, US)

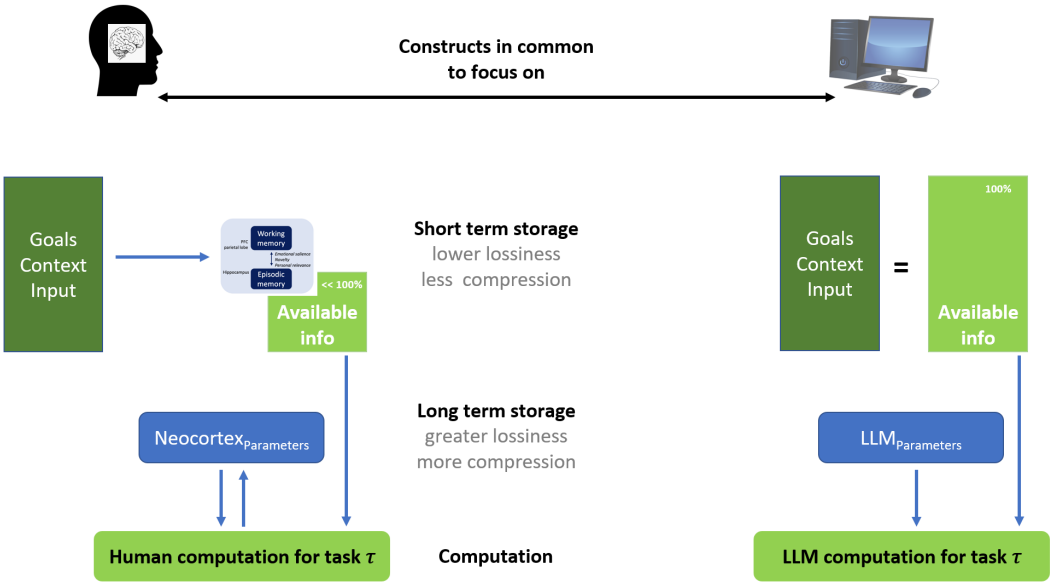
Our group discussed differences and similarities between human linguistic memory and possible LLM analogues. We initially set out to explore a number of themes, including

- Different kinds of memory (working/procedural/semantic/meta-memory)
- Tasks providing insights on memory: summarization, long-form QA, coreference resolution etc.
- Lossiness and information compressed in memory for humans vs. LLMs
- The nature of reasoning and inference, the role of analogy and scratchpads in Chain of Thought generation
- Structured units in representations such as entities/events/common ground and their prioritization/salience
- Implications for Cognitive Science and linguistic theories such of information and discourse structure
- Memory in dialog, including user and conversation memory, dialog states and context representations

We identified a number of problems in the analogy between humans and LLMs, including issues with current RAG-based approaches to long-term memory, which is not well-updated (RAG may retrieve unordered facts, some of which are no longer true or internally contradictory). Some analogies between LLM and human memory concepts are illustrated in the following table:

Brain	Concept	LLM
Experience -> Short term storage	Short Term Memory (STM) - Low compression	Context
Salience bias (emotional, semantic, personal…)	Prioritization at perception time	?
Sleep		(retraining)
Long Term Memory	Long Term Memory (LTM) - High compression	Parameters
Can't know tasks in advance due to linearity of time; must derive all-purpose representations	Task	Can decide at inference time what to attend to based on lossless storage (not lossless *access*, pace Ryan)

After some discussion about how and why humans remember some things (such as how to get back to a weapon store in Trier to catch a shuttle back to Dagstuhl) but not others (what was in the store's window), the group eventually focused on the imperfect human recall of prior linguistic interactions versus LLM access to a potentially perfect transcript in long-context prompts (chat history). Discussion clarified that LLM access to context as memory was mediated in practice by attention bottlenecks at generation time for the model, but this is still not equivalent to human memory, which must prioritize content at perception time, and not at recall time. This disparity is illustrated in the following chart:



The group concluded that some of the most pressing questions are how more human representation types could be incorporated into language technology, though we agreed that evaluation would be a major challenge in establishing both whether certain representations

are in fact more human-like, and whether/how this might be helpful in practical terms. A key direction that arose from discussions in terms of approaching this challenge was the need for constraints on LLM access to prior context, and attention to the characteristics of a system as a coherent agent, for example possessing underlying assumptions and knowledge about oneself. Whereas humans have a notion of "who they are" and therefore of what content they attend to more or less, LLMs are more "generic" and internally inconsistent by nature.

The group continued to communicate after the seminar, including at informal meetings during the following ACL conference in Vienna, and we have set up a mailing list (`lingmem@googlegroups.com`) to keep in touch about the topic. Some members have proposed to collaborate on a potential paper with experiments targeting memory comparisons between humans and LLMs. One proposal was to benchmark humans on a closed and open book summarization task (with and without the text remaining available for consultation after being read once), and comparing LLM behavior with full or partial/attenuated access to the preceding context to simulate a type of compression of information during a left-to-right pass on the text in the closed book scenario. It was suggested that this could be achieved by using existing spoken and written English data with salience scores for sub-spans of input text representing mentioned entities (Ling & Zeldes 2025). A second proposal suggested similar experiments applied to emotional salience, personal relevance or novelty, using data from human dialog annotated for common-ground problems (Sarkar et al. 2025).

Participants

- David Adelani
MILA – Montreal, CA
- Antonios Anastasopoulos
George Mason University – Fairfax, US
- Gašper Beguš
University of California – Berkeley, US
- Verena Blaschke
Ludwig-Maximilians-Universität München, DE
- Ryan Cotterell
ETH Zürich, CH
- Marie-Catherine de Marneffe
UC Louvain-la-Neuve, BE
- Katherine Demuth
Macquarie University – Sydney, AU
- A. Seza Doğruöz
Ghent University, BE
- Robert Frank
Yale University, US
- Juan Luis Gastaldi
ETH Zürich, CH
- Adele Goldberg
Princeton University, US
- Coleman Haley
University of Edinburgh, GB
- Aurelie Herbelot
Denotation – Pritzwalk, DE
- Yu-Yin Hsu
Hong Kong Polytechnic University, CN
- Mark Johnson
Macquarie University – Sydney, AU
- Najoung Kim
Boston University, US
- Alessandro Lenci
University of Pisa, IT
- Lori Levin
Carnegie Mellon University – Pittsburgh, US
- Roger Levy
MIT – Cambridge, US
- Xixian Liao
Barcelona Supercomputing Center, ES
- Kyle Mahowald
University of Texas – Austin, US
- Tom McCoy
Yale University, US
- Joakim Nivre
Uppsala University, SE
- Alexis M. Palmer
University of Colorado – Boulder, US
- Christopher Potts
Stanford University, US
- Jakob Prange
Universität Augsburg, DE
- Siva Reddy
MILA – Montreal, CA & McGill University – Montreal, CA
- Philip Resnik
University of Maryland – College Park, US
- Anna Rogers
IT University of Copenhagen, DK
- Rachel Rudinger
University of Maryland – College Park, US
- Gözde Gül Şahin
Koç University – Istanbul, TR
- Asad Sayeed
University of Gothenburg, SE
- Nathan Schneider
Georgetown University – Washington, DC, US
- Noah A. Smith
University of Washington – Seattle, US
- Mark Steedman
University of Edinburgh, GB
- Tiago Torrent
Federal University of Juiz de Fora, BR
- Bonnie Webber
University of Edinburgh, GB
- Ethan Wilcox
Georgetown University – Washington, DC, US
- Adina Williams
Meta Platforms – New York, US
- Amir Zeldes
Georgetown University – Washington, DC, US

Report from Dagstuhl Seminar 25302

NatureHCI: Towards Designing Computer-Enriched Nature Experiences

Masahiko Inami*[1], Michael Jones*[2], Zhuying Li*[3], Florian 'Floyd' Mueller*[4], and Maria F. Montoya†[5]

1 **University of Tokyo, JP.** inami@inami.info
2 **Brigham Young University – Provo, US.** jones@cs.byu.edu
3 **Southeast University – Nanjing, CN.** zhuying@exertiongameslab.org
4 **Monash University – Melbourne, AU.** floyd@floydmueller.com
5 **Monash University – Melbourne, AU.** maria.montoyavega@monash.edu

Abstract

This report documents the proceedings and outcomes of a NatureHCI seminar, in which 21 researchers and academics from across the world gathered at Schloss Dagstuhl, Germany, to discuss the grand challenges that this field currently faces. We present the activities developed day by day, including PechaKucha self-introductions, collaborative workshops, hands-on design sessions, and group discussions. Finally, we present the pathways that attendees proposed to start addressing the seminar's critical question: how interactive technologies can be designed responsibly to improve our experience of nature, thereby strengthening our connection with nature, which benefits health and wellbeing? Ultimately, with this report, we hope to inspire upcoming Dagstuhl Seminar proposals interested in advancing the field of HCI.

Seminar July 20–25, 2025 – https://www.dagstuhl.de/25302
2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI)
Keywords and phrases NatureHCI, Nature Interactions, Wilderness, Technology, Wellbeing, Sustainability, More-Than-Human, Human-Computer Interaction, non-humans, animals, nature
Digital Object Identifier 10.4230/DagRep.15.7.213

1 Executive Summary

Michael Jones (Brigham Young University – Provo, US)
Masahiko Inami (University of Tokyo, JP)
Zhuying Li (Southeast University – Nanjing, CN)
Florian 'Floyd' Mueller (Monash University – Melbourne, AU)

In July 2025, the seminar brought together 21 international experts from diverse fields, including Human-Computer Interaction, Environmental Science, Design, Psychology, and Cultural Studies, to explore how digital technologies can facilitate and enhance our interactions with the natural environment. The following report documents the seminar and the efforts of the participants to investigate the underlying opportunities and challenges in developing interactive systems to engage with nature.

* Editor / Organizer
† Editorial Assistant / Collector

NatureHCI: Towards Designing Computer-Enriched Nature Experiences, *Dagstuhl Reports*, Vol. 15, Issue 7, pp. 213–252
Editors: Masahiko Inami, Michael Jones, Zhuying Li, Florian 'Floyd' Mueller, and Maria F. Montoya
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

We define "NatureHCI" as research in which the researcher's intent is to learn something about how interactive computing technology mediates or might mediate engagement with nature. This definition intentionally allows for diverse interpretations of "nature" while focusing on the mediating role of technology in human-nature relationships. Interactive technologies offer instrumental benefits by guiding individuals to natural areas, using enhanced navigational aids and disseminating information about these environments through visualization techniques. Additionally, machine learning algorithms can identify patterns in nature-related activities, aiding in planning. On an experiential level, technologies such as virtual reality can simulate natural settings for those with limited access, wearable technologies can enhance sensory experiences, and drones can provide innovative forms of visual interactions within natural landscapes. In response, such NatureHCI systems can support novel engagements based on different cultural and human perspectives.

This body of work shows the growing research interest in NatureHCI systems. Efforts to map the field and to provide systematic ways to design HCI systems have recently emerged. However, what is still missing is a coherent framework for understanding how these technologies support positive and responsible interaction with nature. This must be investigated urgently as a response to a profound planetary shift: The fact that the anthropogenic mass exceeded all living biomass in 2020 demonstrates a fundamental transformation in the relationship between technology and nature. This historic turning point elevates the question of how to design and evaluate technology use in natural environments from academic interest to planetary necessity.

In this context, NatureHCI emerges not as another subdiscipline of HCI but as an essential response to existential challenges. Climate change accelerates ecosystem disruption while biodiversity loss approaches irreversible tipping points. Simultaneously, urbanization disconnects billions from direct nature experience, leading to what Richard Louv termed "nature deficit disorder." Indigenous knowledge systems face erosion just when their wisdom is most needed. Yet this moment of crisis also presents unprecedented opportunity. Ubiquitous computing enables new forms of nature engagement, sensor networks can monitor ecosystem health at previously impossible scales, augmented and virtual reality technologies can bring nature experiences to those physically distant, artificial intelligence begins to decode non-human communication, and global connectivity enables rapid sharing of solutions across cultures and ecosystems.

This seminar builds on more-than-human design theory, post-phenomenology, and environmental psychology, demonstrating that NatureHCI requires fundamentally interdisciplinary approaches that balance technological innovation with ecological sensitivity, cultural awareness, and ethical responsibility. Therefore, this seminar invited experts from around the world and with diverse backgrounds, who could recognise the importance of integrating technological advances with environmental sciences and indigenous knowledge systems to address critical gaps in the field. Most importantly, the diversity of participants allowed for the discussion of several methodological approaches that could begin to constitute a distinctive NatureHCI approach, differing markedly from laboratory-based HCI methods.

The seminar produced several significant outcomes that will shape NatureHCI's development as a field. First, participants developed an initial identification of a grand challenges framework that addresses the complexity of designing technology for natural environments. These challenges emerged from recognizing that HCI's traditional approaches often fail when confronted with nature's unpredictability, multiple timescales, and more-than-human stakeholders. Since then, the elucidation of these grand challenges has been finalized and submitted to a major conference in the field. Perhaps most surprisingly, an animated dis-

cussion on the final day led to a proposal to form a special interest group on "Augmented Animals." This emerged from Professor Inami's observation that "augmented humans are now too small – maybe we can say augmented animals." Rather than augmenting animals for human benefit, this initiative explores how technology might help other species adapt to human-modified environments. Beyond academic outcomes, the seminar catalyzed a vibrant research community committed to collaborative action. Participants are committed to creating an open repository of nature-entangled design methods, making innovative approaches accessible to researchers worldwide.

2 Table of Contents

3 Seminar Goals and Format

The seminar pursued four interconnected goals that shaped its structure and outcomes. First, establishing conceptual foundations for NatureHCI as a coherent research area required defining scope and boundaries, identifying core research questions, mapping relationships to adjacent fields, and developing shared vocabulary. Second, identifying grand challenges provided focus for future research efforts. Rather than generating an exhaustive list of all possible research directions, the seminar aimed to surface the most critical and generative challenges – those that, if addressed, would significantly advance the field while contributing to planetary wellbeing. Third, building community was recognized as essential infrastructure for an emerging field. Academic research often proceeds through individual efforts, but NatureHCI's inherent interdisciplinarity and global scope require collaboration. Fourth, catalyzing concrete action ensured that insights wouldn't remain trapped in academic discourse. Thus, we encouraged participants to create publication plans from collaborative discussions.

Figure 1 Participants engaging in different activities throughout the seminar. Top Left: Sharing coffee breaks. Top Right: Sharing a ride. Bottom Left: Sharing a bus ride to the hike. Bottom Right: Playing volleyball.

To achieve these goals, prior to the seminar, the organizers asked the participants to prepare a pitch (Pecha Kucha format), including the following elements: an introduction of each participant (including their hobbies), a presentation of their relevant NatureHCI

work, their expectations for the seminar, the rationale behind the choice of the recommended reading, and the grand challenges they see in the NatureHCI field. The recommended readings were collected in a shared folder, and more were added throughout the week. A spreadsheet "who has what to offer" was shared prior to the seminar, with a focus on sharing activities in nature that could be done during the week. Participants also engaged in sport activities like running, volleyball, and table tennis. Mornings typically began with provocative presentations or structured workshops, while afternoons allowed for self-organized activities. Some groups conducted impromptu experiments in the castle grounds, testing ideas about technology in nature using available materials. Others engaged in deep philosophical discussions about the nature of nature itself. Evening sessions ranged from formal presentations to storytelling circles where participants shared personal experiences that had drawn them to NatureHCI.

This rhythm – alternating between intensive intellectual work and reflective practice – proved crucial. As one participant observed, "We couldn't have developed these ideas sitting in a windowless conference room. Being surrounded by nature, hearing the birds each morning, walking the forest paths – our environment taught us as much as our discussions."

3.1 Day 1: PechaKucha Self-Introductions and Initial Explorations

The seminar took place in one of the Dagstuhl's conference rooms. Beforehand, the organizers established several work group materials, such as whiteboards, flipover sheets, sticky notes, and markers, to allow participants to record their thoughts and opinions and share them at a glance within the room. The seminar opened with an interactive icebreaker designed to encourage quick connections among participants and cultivate a playful, enthusiastic atmosphere. Participants formed a circle, each introducing themselves by stating their name, accompanied by a distinctive bodily gesture. The group then repeated both the name and gesture. With each new introduction, all previous names and gestures were reiterated in sequence. After the engaging introductory activity, organizer Florian 'Floyd' Mueller provided the opening remarks, which contextualized the seminar and set the basis for the discussion of designing for NatureHCI. Floyd encouraged participants to add, while listening through the various introductions, their thoughts on "challenges and opportunities" and the "taxonomy classifications" of designing for NatureHCI to the available flipover sheets at any time during the session.

Next, each participant introduced themselves through a PechaKucha-format presentation. The constrained format forced presenters to distil their vision for NatureHCI into compelling narratives rather than comprehensive literature reviews. Each participant shared their work related to natureHCI, including prototypes, findings, recommended readings, and grand challenges identified through their work and experience. This approach was inspired by earlier HCI efforts that highlighted key challenges during prior similar seminars. The summary of the presenter's talk is documented in section 4.

3.2 Day 2: Demonstration session and group discussion on the NatureHCI challenges

To spark discussion and inspire creative thinking, several participants brought demos of their NatureHCI work. During a dedicated two-hour session, participants explored eight interactive demo tables. These ranged from water-based interactions through surfing, to plant-responsive interfaces, footwear interactions that adapt music based on walking or

Figure 2 Pecha Kucha sessions, where all participants introduced their research related to SportsHCI.

running, a navigation glove for skiing in the mountains and an artifact designed for bird sound identification. Presenters shared their motivation and inspiration behind each demo, explained how they designed their systems, and reflected on their key learnings, highlighting what worked, what could be improved, and what they might approach differently in future iterations. These demonstrations are detailed in section 5.

Figure 3 Group discussions carried out during the seminar.

The remainder of the day was spent working further on elaborating Grand Challenges for the field of NatureHCI. This was prepared beforehand by organiser Michael Jones through a collaborative brainstorming session. Professor Jones began the brainstorming activity by using the following three questions as prompts: What is NatureHCI? What is important to NatureHCI? And what are the next problems to solve in NatureHCI?

To explore these questions, participants were divided into smaller groups of four to five members. Each group appointed a scribe to document ideas and a reporter to present the group's discussion insights. The brainstorming activity was structured into three rounds, each corresponding to one of the guiding questions. For each round, groups were given 10 minutes to discuss the question and record their thoughts on one-third of a shared flip chart designated for that round. Following the group work, each reporter presented their group's reflections in a 5-minute report, which was then followed by a brief open discussion. This

opened up the discussion to the question of what nature is. What is the difference between NatureHCI, OutdoorHCI and SportsHCI? How can we define the boundaries between reality and virtuality? What design tools and methodologies are needed to support research and practice in NatureHCI?

3.3 Day 3: Reflections on the NatureHCI domain and contributions; Hike

On the morning of the third day, organiser Michael Jones continued to guide the discussion on the previous day's main ideas. Participants were reshuffled and reorganized into three new groups to further explore key thematic questions. Each group focused on one of the following: (a) What is NatureHCI? (b) What design methods and adaptations are needed for NatureHCI? and (c) What are the boundaries between NatureHCI, SportsHCI, and OutdoorHCI? The third question was explored through the use of a Venn diagram, with participants tagging examples and ideas into the relevant categories to better understand where overlaps and distinctions emerged. The boundary cases identified during the Venn diagram activity were later revisited in a one-hour plenary session with the full group of experts. This session provided an opportunity to further unpack and discuss ambiguous or overlapping examples, aiming to reach a clearer understanding of how these cases fit within the categories.

Figure 4 Participant discussing NatureHCI's scope with different examples from SportsHCI and OutdoorsHCI.

In the afternoon, participants joined the traditional Saarschleife Tafeltour hike in the region. Organizers suggested taking the opportunity of walking together to share career mentoring advice. The organizers proposed a walk at the treetop lookout tower at Saarschleife while engaging in a discussion about whether the tower was disrupting the natural landscape or nicely integrated. The participants observed and discussed the treetop structure in relation to the seminar topics and how they navigated the hike: First, they noticed the interactive elements spotted at the site directed their thinking towards stakeholder experiences and involved them in the design process. For example, tree boxes for squirrels (including an image of a squirrel using the tree box) should factor in children's accessibility when designing interactive and informative installations. Second, participants' attention was directed to the dominant species, events, and challenges faced at the ecological site through installations.

Third, most participants disagreed with the need for the tree-top structure, citing the disruption of the natural landscape and the existing natural lookout as their reasons.

Figure 5 Participants engaging in the traditional hike.

3.4 Day 4: Grand Challenges identification

In the morning of the fourth day, organizer Masahiko Inami facilitated a strategic discussion on ways to group the grand challenges identified so far. The focus was now on identifying in the notes both what is currently known and what remains uncertain about NatureHCI and designing for interactions with and within natural environments. These reflections provided a foundation for collaboratively identifying the grand challenges in the field. A "grand challenge" was defined as an important and difficult problem that usually requires long-term effort, often spanning a decade, and has the potential to inform multiple research questions. Each grand challenge was framed in terms of a "lack of knowledge" or "limited understanding", and included an implicit or explicit call to action for the research community. We clustered the collected data in a collaborative and reflexive manner, where discussions evolved from practice and theory in an intertwined way, going back and forth between design examples and abstract knowledge. As we wanted to let the participants' experiences drive the emergence of the individual challenges, we allowed the clustering of the notes from the initial presentations to emerge organically. Finally, participants were divided into groups to elaborate on proposed themes, including design-related grand challenges, human–nature relationship challenges, technology-related challenges, and those concerning users and stakeholders. Midway through the activity, the groups were reshuffled to encourage the exchange of diverse perspectives. The groups presented the highlights in a plenary session, providing a brief description of each possible revised challenge and its associated sub-challenges. We then iteratively refined these to create our final grouping of grand challenges, aiming to reach consensus while capturing the key ideas discussed. In Section 6, we present the identified grand challenges in detail.

3.5 Day 5 Wrap-up, Conclusive Remarks, and Concrete Follow up Actions

The last morning of the seminar was devoted to wrapping up all the ideas and insights collected during the week. The organisers proposed a dedicated writing session to advance on the first collaborative outcome of the seminar, a Grand Challenges paper for a high-level HCI venue.

By the end of that morning, the organizers were mainly dedicated to the report, but also defined and divided the responsibilities for the follow-up actions. For all concrete "next step actions", and based on personal interests and preferences, the group assigned a main person as the "lead" and listed all the participants interested in committing to work on the task.

4 Overview of Talks

4.1 Human Augmentation in Natural Environments

Masahiko Inami (University of Tokio, JP)

Drawing inspiration from 17th-century scientist Robert Hooke's conception of the microscope as an "artificial organ" to enhance the five senses, Inami positioned technology not as separate from nature but as a means to deepen our relationship with it. His proposed "Supersensory Nature Perception" aims to make perceivable natural phenomena beyond human perception – such as UV patterns, infrasound, and magnetic fields – thereby fostering greater empathy and understanding of the natural world.

Practical implementations include the "MagniFinger" finger-mounted microscope system designed to transfer expert botanists' skills to beginners, technology that visualizes plant-insect interactions in real-time, and gut microbiome sensing research that deepens understanding of internal natural ecosystems. These innovations embody an approach that incorporates Japanese concepts of wa (harmony) and ma (space/pause), extending human sensory capabilities while respecting natural rhythms and cycles rather than seeking to control nature.

My research transcends Western dualism of nature versus technology, suggesting possibilities for constructing new relationships between humans and nature through augmentation technology. Through interdisciplinary collaboration with botanists and ecologists, his work integrates technological and ecological perspectives, facilitating new discoveries and understanding in the natural world.

4.2 Exertion and nature

Florian 'Floyd' Mueller (Monash University – Melbourne, AU)

Exertion experiences can benefit from nature, and interactive technology can support this. We demonstrate this through a series of research design works around integrated nature exertion experiences. The results of these works suggest interesting ways forward for NatureHCI

research, in particular, how the design can highlight experiential aspects, facilitating playful experiences. Ultimately, with our work, we want to enhance our knowledge around the design of NatureHCI experiences to help people understand who they are, who they want to become, and how to get there.

4.3 Cultural Perspectives on Nature and Technology

Michael Jone (Brigham Young University – Provo, US)

Outdoor recreation creates opportunities for people to engage with nature. I think of engagement with nature as engagement the specific time, place, and communities in which the recreation takes place. For example, engagement with time includes engagement with time-varying conditions that exist during the recreation experience. Framed this way, we can talk about interactive computing technology use during nature recreation in terms of it's impact on engagement with time, place, and community. I take a post-phenomenological view on how technology impacts engagement. In the post-phenomenological view, technology mediates engagement with nature to create different forms of engagement. Taking the post-phenomenological perspective on engagement with place, time and communities creates a rich design space in which we can explore different ways in which different technologies create new forms of nature recreation.

4.4 Technological Mediation of Human–Nature Relationships

Zhuying Li (Southeast University – Nanjing, CN)

Human-nature interaction contributes significantly to wellbeing. Within the context of NatureHCI, I am interested in how technology can mediate the relationship between humans and nature, and how people experience nature, particularly in urban environments where direct contact with natural elements has become increasingly limited. This inquiry can be approached from various dimensions. One important aspect is the spatial-temporal dimension. Due to constraints of space and time, people often have limited opportunities to engage with nature. Technology can help identify and create such moments of engagement, supporting more frequent and meaningful interactions. Another dimension concerns our lived experience in highly digital and urban settings, where human attention is often directed toward screens and mediated activities. Here, technology can play a role in redirecting attention back toward natural phenomena and fostering a deeper sense of awareness and connection. Finally, I would like to extend this exploration toward accessibility and inclusivity, ensuring that technologies designed to reconnect humans with nature can benefit diverse groups and abilities. Through these perspectives, I hope to understand how NatureHCI can support equitable, restorative, and sustained forms of human-nature engagement that enhance overall wellbeing.

4.5 Virtual Nature

Tuomas Kari (Natural Resources Institute Finland (Luke) – Helsinki, FI

As an orienteer from childhood and current active trail runner, I have ventured in nature all my life. I have always been interested in technology, nature, and wellbeing. My current research topic, virtual nature, combines these three areas of personal and professional interest. Virtual nature refers to solutions where a nature environment is delivered virtually by utilizing digital technology. Research has shown that virtual nature can induce various psychophysiological benefits to its users. Virtual nature can be used as a recovery space in workplaces, educational facilities, and similar, as well as to bring nature experiences to people who have limited access to nature, such as people with disabilities, aged people in elderly care homes, hospital patients, people living in large cities, etc. A central challenge in disseminating virtual nature solutions more widely is to make (more) people understand that virtual nature is not meant to replace actual nature contact but rather to provide an alternative in situations where (and when) access to actual nature is limited or not available. It is also central to research and develop virtual nature solutions to meet different purposes.

4.6 Local Perspectives and Cultural Heritage in NatureHCI

Siiri Paananen (University of Lapland – Rovaniemi, FI)

Nature is sometimes perceived as untouched or empty, but landscapes are woven with stories, meanings, and histories – especially when viewed through local perspectives. My work in NatureHCI, at the Lapland User Experience Research Group, explores how interactive technologies can reveal and highlight these hidden layers, making them accessible and engaging for wider audiences. I'm interested in how digital tools like XR and AI, together with participatory design, can support a pluriversal approach – acknowledging and valuing different relationships with nature, especially those from Indigenous and local communities. For me, it's important that NatureHCI works alongside these communities, using technology in respectful and meaningful ways. At Dagstuhl, I look forward to exploring how we can design technology that supports connection to the living landscapes we share. This work builds on our research group's experience in NatureHCI, as well as combining outdoor and cultural heritage HCI research.

4.7 Challenges and opportunities for nature-oriented design

Sarah Webber (University of Melbourne, AU

As a lifelong camper and hiker, I use a range of mobile apps to explore and learn about wildlife and landscapes. This reflects a scientific nature learning orientation. However, people adopt varied modes of nature engagement, and so can benefit from different types of

digital tools, and have different needs and tolerances for distraction, unobtrusiveness, and mediation. More-than-human thinking challenges HCI researchers to develop design methods that can include non-humans and ecosystems as stakeholders. Amidst mounting concern regarding e-waste, environmental degradation and slowing of human population growth, new priorities emerge, such as decentering the human, and designing systems that support ongoing familiarity and relationships with non-humans and ecosystems. Models of post-growth innovation, zero-waste products, and systems to support communities in managing commons, may be fruitful directions for HCI to contribute to flourishing socio-environmental systems.

4.8 Reimagining Nature HCI: Connecting Inner and Outer Ecologies Through the Sensory Body

Nandini Pasumarthy (Monash University – Melbourne, AU

Our relationships with nature are not just visual or cognitive – they are embodied, interoceptive, and shaped by inner ecologies like breath, gut rhythms, and sensory memory. Drawing on research in gut health, play, and interaction design, this abstract explores how NatureHCI might move beyond representation to attune with what the body already senses. It proposes a design approach grounded in energetic coherence and local co-creation, where user interaction with nature aligns with the rhythms of place, season, and sensation. Through playful, multisensory, and body-led experiences, this abstract reimagines NatureHCI as a space for co-regulation, grounding, and reciprocal connection between inner and outer ecologies.

4.9 How can HCI engage people from kindergarten to old age in caring for biodiversity?

Margot Brereton (Queensland University of Technology – Brisbane, AU)

Habitat loss, invasive species and climate change are significantly affecting biodiversity and ecosystems, causing changes to the abundance and geographic range of many species, interfering with their life cycles and interactions with other species. Although continental scale sensing infrastructure to monitor biodiversity is in place, and AI has begun to classify fauna sounds, technology needs significant help from people to make accurate findings. Moreover, data collection and scientific analysis alone do not build broader community interest and capability to act. Current approaches to citizen science focus on scientific need rather than building user interest (e.g. Zooniverse), offering top-down, repetitive, highly structured tasks to classify data e.g., is this frog call present in this sound bite? This talk discusses the need for new approaches to engage people with nature. From local concerns and community members as young as those in kindergarten, we discuss how HCI can scale from local to global to develop interest in and custodianship of our natural environment and all of its amazing species.

4.10 Why should NatureHCI investigate outdoor water activities?

Maria Fernanda Montoya Vega (Monash University – Melbourne, AU)

Aquatic environments encompass 70% of the natural environments on the planet, including the ocean, lakes, rivers and other freshwater sources. Human engagements with such natural environments span many industrial, cultural and recreational activities, approaching water as a resource or as a site for leisure. However, within HCI, recreational engagements with aquatic environments have not yet fully benefited from interactive technologies beyond limited applications focusing on athletic performance. Similar to how interactive technologies have enhanced the enjoyment of on-land outdoor activities, I see the potential of interactive experiences to boost IN-SITU experiences in aquatic environments, which could support their conservation, raise ecological awareness and promote physical and mental health. Hence, given the scale and importance of aquatic ecosystems and the potential for interaction design to shape the future interactions in such environments, this domain deserves focused attention within NatureHCI.

4.11 Accessibility in NatureHCI

Jasson Wiese (University of Utah – Salt Lake City, US)

Many outdoor experiences are inaccessible by default. They were not necessarily designed that way–the natural world can be a harsh and unforgiving place. How then should we think about accessibility when we are designing technological experiences for the outdoors? Do we make our designs accessible even if the environment in which they are used is not accessible? Do we focus on trying to make nature more accessible? Or do we simply ignore accessibility concerns in this context because it's simply too hard? I'd like to challenge this community to engage meaningfully with accessibility as a first class consideration for the outdoor contexts in which we are working. Finally, if we do accept this challenge we need to engage with how we can ensure participation by all people, not simply the "least disabled," who are typically the easiest to engage with traditional HCI methods.

4.12 Human-Plant interaction

Hong Luo (Monash University – Melbourne, AU)

Advances in interactive technologies offer new opportunities to enhance human–nature engagement. Although human–plant interaction (HPI) research has begun to explore this area, current systems appear to focus on unidirectional interactions between plants and humans. In contrast, the idea of bidirectional interactions where technology mediates a mutual exchange between humans and plants has received far less attention. As a new form

of interaction between humans and nature, bidirectional interactions provide people with opportunities to re–engage and experience nature, while offering the potential to enhance the connection between humans and nature. Given the significance of fostering such reciprocal relationships, this area of research warrants greater focus, with the aim of enriching our bond with nature and moving toward a more integrated coexistence.

4.13 Augmenting human experience and skills in mountainous outdoor environments

Florian Daiber (DFKI – Saarbrücken, DE)

In the last years, sports technology has become ubiquitous and there has been a large body of work in HCI as well as commercial products including apps and wearables that aim to enhance human experiences in the outdoors. Moving in (extreme) outdoor environments requires certain physical and cognitive skills and HCI in the outdoors has the potential to track human experiences and provide guidance to exist in and with nature respectfully. In our research we investigate the adoption and usage of technologies in outdoor activities such as mountaineering, climbing and running with a particular focus on their interaction with nature. We are especially interested in technologies that track and assist users in their individual needs in nature. The preparation for as well as assistance during the outdoor activity is of interest. In [18]we investigated how climbing technology that enables similar features as running and cycling technologies is perceived by climbers. The goal of the survey is to gain insight in the acceptance of technology in climbing. The main finding of the survey is that the sample can be divided into leisure-oriented outdoor climbers and sports-oriented indoor training enthusiasts. In [16] we investigated how technology intersects with the concept of mastery in the context of mountaineering, a domain that deeply values skill, autonomy, and experience. Rather than focusing solely on usability or efficiency, we argue that designers must consider how technological interventions impact the experiential, ethical, and cultural dimensions of outdoor sports in nature.

4.14 Sensory immersion in NatureHCI

Carey Jewitt (University College London, GB)

There is a need and benefits to fostering alternative visions of digital communication. The rapid expansion of techniques of simulation and sensory manipulation beyond the visual in virtual and mixed reality is central to the futures trajectory of digital communication. This sensory turn is primarily predicated on high-fidelity human-centric replication of complex human sensory processing. However, the dominance of a human-centric replication paradigm can be problematic and limiting for HCI. 1) It grounds digital communication in a technical and limited view of the senses which severely constrains the range of digital sensory experiences available to people and can lead to misaligned sensorial design. 2) Tends to presuppose a homogeneous, able-bodied user, thereby excluding countless other embodiments

and possibilities of communication. 3) Locks us in our current ways of perceiving and knowing and restrictive modes of relating to one another. 4) Holds back innovation by anchoring the digital in the physical human world. Drawing inspiration from animal and plant sensory worlds provides a speculative springboard to generate new reconfigurations for sensorial immersion technologies beyond the human centric.

4.15 Designing With, Not Just For, Nature: A Post-Anthropomorphic Approach to Human-Plant Interaction

Rakesh Patibanda (Monash University – Melbourne, AU)

This talk traces an evolving relationship with plants, from early life experiences to advanced research in mental health, virtual reality, and bodily interaction. This trajectory culminated in PlantMate, a system exploring shared bodily control between humans and plants through playful engagement. Initial work, focused on technological novelty, prompted critical reflection on embedded anthropocentric biases in design. This introspection led to a reframing of interaction design through a post-anthropomorphic lens. Drawing on recent advances in more-than-human design, I presented four artifacts: Ant Apparatus, Breathing with Plants, Being with Plants, and Symbiotic Breather. These projects serve as provocations, urging a reconsideration of technology not merely as a human tool, but as a mediator in entangled, multispecies relationships. I conclude by posing open questions about designing with nature, emphasising the importance of listening across differences, embracing ambiguity, and enabling nonhuman actors to co-author shared futures.

4.16 Tech on the Trail: Investigating Nature on Socio-technical Journeys

Scott McCrickard (Virginia Tech – Blacksburg, US)

This work seeks to understand and develop ways that technology is used (or avoided) on trails and in trail-like settings, especially extended and multi-day hikes, where different user goals and desires affect our behaviors and interactions with others. Technology is often targeted for use in heavily populated urban environments, but thousands of people take technology on their outdoor adventures, raising questions about appropriate use when in a more isolated and natural environment. These environments provide some level of separation for most people from technologies, but a need for community and communication still exists for hikers and different groups and trail stakeholders (e.g., friends on the trail, family back home, trail maintainers) and on society.

4.17 Real-World Terrain and Gait Recognition with Ground-Aware Smart Footwear

Don Samitha (Monash University – Melbourne, AU)

Everyday, billions of people use footwear for walking, running, or exercise. Of emerging interest are "smart footwear", which help users track gait, count steps or even analyse performance. However, such nascent footwear lack fine-grain ground surface context awareness, which could allow them to adapt to the conditions and create usable functions and experiences. Hence, this research aims to recognize the walking surface using a radar sensor embedded in a shoe, enabling ground context-awareness. Using data collected from 23 participants from an in-the-wild setting, we developed several classification models. We show that our model can detect five common terrain types with an accuracy of 80.0% and further ten terrain types with an accuracy of 66.3%, while moving. Importantly, it can detect the gait motion types such as "walking", "stepping up", "stepping down", "still", with an accuracy of 90%. Finally, we present potential use cases and insights for future work based on such ground-aware smart shoes.

4.18 Enabling Mobile XR Interaction in Dynamic Natural Environments

Andrii Matviienko (KTH Royal Institute of Technology – Stockholm, SE)

Interaction on-the-go allows users to interact with technology in dynamic contexts, e.g., walking, running, driving, or flying. The biggest challenges with this type of interaction are (1) users' inattention to surrounding environments, making traditional forms of input difficult to use, and (2) users' inability to use their bodies freely due to interactions that require certain body parts. Particularly, interaction on-the-go in XR is largely underexplored in nature and for interaction with and within nature and requires special attention and consideration since people in nature engage with different and varying activities, spanning from hiking and climbing to camping and playing games. XR enables interactions and visualizations that can represent objects that are occluded or are impossible to see with biological vision. Moreover, existing XR experiences and solutions are typically limited to individual, indoor, visual, and auditory perceptions, neglecting the outdoor environment in nature.

4.19 From Story to Shift: Designing Climate Empathy Through Games and Behavioral Insight

Ambika Shahu (IT:U Interdisciplinary Transformation University – Linz, AT)

What if experiencing the life of a farmer affected by climate change could change how we think about our own actions? Through empathy-driven role-playing games, players must make tough decisions as a farmer grappling with climate change, choosing between short-term survival and long-term sustainability. This interactive storytelling approach draws players into complex scenarios, building emotional connections and prompting reflection on real-world climate choices. Alongside, we also explore how psychological factors such as habits, perceived control and social expectations influence people's willingness to adopt greener commuting methods, such as using public transport, cycling or shared automated vehicles. Our research reveals how mental shortcuts, policy gaps and the perception that climate change is "too far away" (in terms of time, space or relevance) can hinder action. By blending narrative with behavioural science, we aim to inspire new approaches to designing technology that brings climate realities closer and sparks meaningful change.

4.20 Accessibility in the nature

Amad Alsaleem (University of Utah – Salt Lake City, US)

When we talk about accessibility and nature, the conversation usually centers on ramps, paths, and removing obstacles to access. But the real barrier may lie in how we design experiences, especially we aim for safety only not the possibility. In our work, through the design of adaptive shared-control systems, we empower individuals with tetraplegia to engage in extreme sports like sailing and skiing, not as passengers, but as active participants. Based on our work, I will talk about how adaptive systems can distribute control while leaving critical decisions in the hands of the user. The result: users gain agency, not just access – by designing experiences that give users the right to fail in a safe and meaningful way. This talk will discuss the design challenges of shared-control interfaces in unpredictable natural environments and explore how technology can support fuller, more enjoyable outdoor experiences.

4.21 Bringing social sciences to the management of nature and outdoor experience

Bill Borri (Deakin University – Melbourne, AU)

I travel from the lands of the Wurundjeri and the Salish peoples. As a conservation social scientist, I bring both quantitative and qualitative research methods. I am a pluralist, with various epistemological, ontological and axiological stances. Q. What does it mean to "be"

in Nature? At the University of Montana, I focused on wilderness and wildlands: a place apart, a freedom to be, beyond the hand of human domination. The USA shares a heritage from nature, with a geography of hope. Q. What types of place, identity and character? Many Americans and Australians have begun and grown their relationship to nature on public lands and, yet, we live in a time of increasing privatisation and private control of nature. Disney, as one of the largest managers of outdoor experiences, creates mediated, commodified, and constructed nature. Q. Will hyper-reality overwhelm the real?

5 Overview of interactivity session

The second day of the seminar was introduced by organizer Michael Jones, starting off with a "demonstration session" that involved live interactive demonstrations of systems and technologies relating to NatureHCI.

Florian Daiber demonstrated FootStriker [19], a self-contained wearable that detects heel striking while running with a pressure-sensitive insole. Heel striking is corrected in real-time to mid/forefoot running by applying electrical muscle stimulation (EMS) on the calf muscle. Florian discussed potential scenarios for EMS-based training in nature with the participants.

Figure 6 Left: Florian Daiber demonstrating the Footstriker. Right: Sirii Paananen demonstrating the Glove Navigator.

Sirii Paananen demonstrated Glove Navigator [23], a wearable graphical navigator display integrated into a skiing glove. Sirii showcased this prototype for participants, which was made using a standard faux leather mitten and a smartphone. A 50x50 mm opening was cut on the mitten's upper part near the fingertips, and the phone was slid into the modified seam. The phone has a navigation app installed, displaying an arrow towards a target. The navigation target can be set up before skiing into the slopes.

Nathalie demonstrated the Virtual Campfire project [8], asynchronous video storytelling using the Marco Polo app, aiming to connect older and younger adults while enhancing engagement with nature. Nathalie shared the study results with participants, highlighting how the project created discussions on personal experiences and memories tied to natural settings. In this project, they found that storytelling served as a useful icebreaker, with nature acting as a catalyst for meaningful intergenerational communication.

Hong Luo presented PlantMate [32] a platform enabling bidirectional touch-based interaction. PlantMate translates users' touch into bioelectrical stimulation to enhance plant growth while translating a plant's electrical signals under varying environmental conditions

(e.g., temperature, humidity) into electrical muscle stimulation for users. Hong shared with participants how plants can be perceived as interactive agents, redefining human-plant relationships through discernment and affective touch. Participants engaged with this prototype by attaching the EMS electrodes to their forearms.

Figure 7 Left: Nathalie presenting Virtual campfire to Professor Masahiko. Right: PlantMate system being experienced by a participant.

Maria Montoya demonstrated OceanEcho [33], a surfing wearable embedded with vibrotactile actuators, a thermal pad, headphones and a location system, aiming to foster surfers' connectedness with the ocean. Participants engaged with OceanEcho by wearing it near the water fountain, simulating ocean contact, while the prototype played nature-related sounds (such as dolphin and whale sounds), providing heat and vibration stimuli that represented the ocean state (swell direction).

Figure 8 Left: Maria Montoya demonstrating OceanEcho to participants. Right: Participant experiencing OceanEcho near the fountain.

Margot demonstrated Ambient Birdhouse [50] a novel IoT design for the home that seeks to encourage awareness and discovery of birds outside. The Ambient Birdhouse can be placed inside a house and plays media of local birds – sometimes giving clues about them. Bird houses are connected and users can share bird media from their phone, challenging other users to identify them. Participants enjoyed listening dto ifferent bird sounds that transported them to nature. Margaot shared with participants that this prototype's playful nature had an immediate grasp on children, prompting them to learn bird calls.

Figure 9 Left: Margot Brereton demonstrating Ambient Birdhouse. Right: Participants engaging and sharing during the demo session.

6 Defining NatureHCI: Scope and Boundaries

The first major topic during day 3 tackled the foundational question that had haunted previous gatherings: What exactly is NatureHCI? Previous workshops had operated with implicit understandings that often diverged significantly. Some participants assumed NatureHCI meant any technology use outdoors. Others focused on environmental sensing. Still others emphasized nature education or conservation applications. This diversity had generated creative work but hindered field formation. The workshop began with individual reflection. Participants spent ten minutes writing about what had brought them to NatureHCI – personal experiences, research questions, societal concerns. Sharing these stories revealed profound differences. One researcher described childhood summers with grandparents who taught plant identification, lamenting how urbanization severed such connections. Another confessed to discovering nature only through Pokemon Go, which led to actual hiking. An indigenous scholar explained how their community saw no separation between nature and technology – both were gifts requiring respectful use.

Small groups then tackled the question "What is and isn't NatureHCI?" through concrete examples. Heated debates arose. Was a fitness tracker used while hiking NatureHCI? Most said no – unless the research investigated how tracking affected nature experience. What about a VR relaxation app featuring forest scenes? Maybe – if it studied whether virtual nature fostered real nature connection. Agricultural sensing networks? Depends – precision farming for maximum yield seemed different from systems helping smallholders work with natural cycles. Through affinity diagramming, key concepts emerged. "Mediation" appeared

repeatedly – NatureHCI seemed less about specific technologies than about how technologies mediate human-nature relationships. "Intent" proved crucial – commercial products might use identical technologies but with different purposes. "Nature" itself resisted definition – participants agreed that prescribing what counts as nature would exclude important work.

The full group then engaged in collaborative wordsmithing, projecting draft definitions and editing in real time. Tensions arose between precision and inclusivity. Computer scientists pushed for specificity while anthropologists warned against premature closure. After ninety minutes of intensive discussion, consensus emerged around the definition that would guide the week: "NatureHCI research is research in which the researcher's intent is to learn something about how interactive computing technology mediates or might mediate engagement with nature."

Participants appreciated how this definition provided clarity while maintaining flexibility. It distinguished NatureHCI from adjacent fields – not all OutdoorHCI was NatureHCI, not all SustainabilityHCI was NatureHCI – while allowing diverse approaches. The focus on mediation opened theoretical connections to philosophy of technology. Leaving "nature" undefined acknowledged cultural differences while requiring researchers to be explicit about their own definitions. The workshop concluded by identifying boundary objects – concepts that would help distinguish NatureHCI work. These included focus on human-nature relationships rather than purely human or purely natural phenomena, attention to how technology transforms these relationships rather than just enabling them, and commitment to understanding rather than just building. With definitional clarity achieved, the seminar could move to identifying specific challenges.

7 Defining the Grand Challenges on NatureHCI

During the Pecha Kucha presentations and group discussions, notes on challenges and opportunities were written on flipcharts for later analysis. After the day 2 brainstorming session, discussing the NatureHCI scope and the next problems to solve in NatureHCI an initial list of grand challenges were identified by participants, which included: access to nature, culture influences on the experience of nature, safety and risk in nature, simulated natural environments, awareness of non-human entities, antropomorphism of nature, variety of scales and live forms in nature, community knowledge of nature, sustainability design for nature, unpredictability of nature, mutualistic experiences in nature, understanding nature and learning from nature, design methods from other disciplines applicable to natureHCI, collaboration with goverment and local communities, connectivity in nature and diversity of users in nature. During day 3, the list was further refined with the following questions considered:

- What is NatureHCI?
- How does NatureHCI overlap with SportsHCI and OutdoorsHCI?
- What design challenges are related to methods and adaptations that are needed for NatureHCI?
- Can the challenge be solved within the next 10 years?

After a general consensus on the definition of NatureHCI, the other questions were the focus of the discussions during day 3. By the end of this day, organizer Michael Joenes documented the reflection of day 3 group discussions on a shared document proposing a final list of grand challenges, including design-related grand challenges, human–nature relationship challenges, technology-related challenges, and those concerning users and stakeholders. These

challenges were also listed using a Miro board, in which we clustered the collected data on the flipover sheets in a collaborative and reflexive manner, where discussions evolved from practice and theory in an intertwined way, going back and forth between design examples and abstract knowledge. As we wanted the participants' experiences to drive the emergence of individual challenges, we allowed the clustering of notes to emerge organically. Overall, high-priority design challenges included designing for non-human stakeholders, addressing temporal mismatches, ensuring equitable access, and developing appropriate evaluation methods. Other challenges were identified in debates around power and connectivity, indigenous and community knowledge, and evaluation metrics.

During day 4, organiser Inami led the discussion on each challenge by creating breakout groups to further discuss and document the proposed four grand challenges themes. Midway through the activity, the groups were reshuffled to encourage the exchange of diverse perspectives. The groups presented the highlights in a plenary session, providing a brief description of each possible revised challenge and its associated sub-challenges. We then iteratively refined these to create our final grouping of grand challenges, aiming to reach consensus while capturing the key ideas discussed. Throughout this process, we remained mindful of the time constraints of the workshop format. In the next subsections, we present a summary of such challenges at participants conceptualized during the final day of the seminar.

On the final day, organizer Michael Jones called for an overview of the challenges discussed. Rather than minor editing, participants undertook wholesale reconsideration of how to frame NatureHCI's challenges. The original structure organized challenges by discipline – technology challenges for engineers, design challenges for designers, and social challenges for social scientists. But NatureHCI's fundamental insight was that such divisions failed when addressing human-nature relationships. A sensing network designed by engineers without considering cultural meanings would fail. A beautifully designed interface ignoring technical constraints would never deploy. Social interventions ignoring material realities would remain theoretical. Through rapid iterative discussion, a new organizing principle emerged. Each grand challenge should focus on a core aspect of human-nature relationships, with implications for different stakeholders embedded within rather than separated out. This structure better reflected the field's commitment to integration and avoided reinforcing disciplinary silos. In the next subsection we show the efforts of participants during day 5 to conceptualized a comprenhensve list of grand challenges in NatureHCI.

7.1 The Challenge of Designing for Non-Human Stakeholders

Perhaps no challenge better exemplifies NatureHCI's paradigm shift than the question of how to design with and for non-human entities. Traditional HCI assumes human users whose needs can be articulated through interviews, observations, and participatory design. But how do we understand the needs of a forest, the preferences of migrating birds, or the wellbeing of soil microbiomes? Thus, there is a lack of understanding on how to design for non human stakeholders.

We recognize that current HCI methods embed anthropocentric assumptions. User-centered design centers humans by definition. Even when we design for pets or livestock, we typically prioritize human interpretations of animal needs. NatureHCI requires more radical approaches that grant genuine stakeholder status to non-human entities.

Recent work provides starting points. Researchers like Tomico and colleagues [57] have proposed methods for engaging with nature as an active participant rather than passive context. Their repository of "nature-entangled design" methods includes techniques like sensory walks where designers attempt to perceive environments from non-human perspectives, temporal mapping that reveals natural rhythms typically ignored in human-centered timescales, and material dialogues where natural elements participate in form-giving. Yet these methods remain nascent, requiring significant development to address NatureHCI's scope.

An associated design challenge is considering the the design for different temporal scales in nature. An oak tree experiences the world across centuries, while mayflies complete entire lifecycles in hours. Soil formation occurs over millennia, while fungal networks communicate in seconds. Thus, it is still unclear how to develop methods sensitive to this temporal diversity. Long-Term Ecological Research provides one model, with studies spanning decades to capture slow environmental changes. But NatureHCI needs approaches that work within typical project timescales while remaining sensitive to longer rhythms. This might involve designing studies that explicitly plan for handoffs across researcher generations, creating protocols for community-maintained observations, or developing simulation methods that compress long timescales into manageable periods.

The challenge extends to entities beyond human sensory capabilities. Many animals perceive ultraviolet patterns invisible to us. Plants communicate through chemical signals we cannot smell. Electromagnetic fields guide migration in ways we barely understand. NatureHCI must develop methods for engaging with these hidden dimensions of nature. This might require technological mediation, for example using sensors to translate ultrasonic bat calls into human-audible frequencies or chemical signals into visual patterns. But such translation raises questions about representation and meaning. How do we ensure translations preserve rather than distort non-human experiences? How do we avoid imposing human interpretive frameworks on fundamentally different ways of being?

Working iteratively with non-human stakeholders presents additional challenges. In human-centered design, we prototype, test, gather feedback, and refine. But how does a tree provide feedback on a design? How do we know if our interventions benefit or harm ecosystem health? NatureHCI needs new protocols for iterative design with non-verbal, non-human participants. In this regard, we can learn from other disciplines. For example, some researchers explore behavioral indicators: does wildlife approach or avoid installations [37, 35]? Do plants thrive or struggle in proximity to technologies[34, 15]? Others investigate physiological measures such as stress hormones in animals, growth rates in plants, and diversity indices in ecosystems [49]. Still others pursue more speculative approaches such as attempting interspecies communication through shared substrates like mycelial networks or water flows [39].

In summary, NatureHCI researchers need to identify appropriate constructs and measures to design for non-human wellbeing, which may require deep collaboration with ecologists, ethologists, and conservation biologists. To start advancing this challenge, researchers can investigate how to recognise risks and potential benefits of technology to non-humans, how to develop tools to navigate and represent knowledges about specific places, flora, and fauna and how to develop guidance to help designers select between, combine, and effectively deploy a range of methods to investigate the needs of non-humans.

7.2 The Challenge of Technology Innovation for Nature Interactions

We discussed the core challenges of developing NatureHCI systems, technologies that could coexist with, rather than resist, natural environments. Instead of assuming that existing devices could simply be adapted for outdoor use, we exposed how current technological infrastructures fundamentally fail in nature. When mapping breakdowns of conventional technologies in the wild, we realised that electronics designed for controlled indoor environments succumbed to water infiltration, temperature fluctuations, and UV degradation. Technological interventions have also shown how plastics cracked, metal connections loosened, and animals damaged or displaced equipment during in the wild studies. Additionally, biological growth, from algae to fungi, tend to obscure sensors and solar panels. These failures reveal an implicit anthropocentric assumption: that nature is a hostile backdrop to be resisted, rather than a collaborator in design. These challenges became the foundation to reflect on a set of development principles for nature-robust systems. Devices must "breathe" without leaking, withstand temperature extremes without active control, and either repel or harmonize with biological growth. They must also be resilient to animal curiosity – coexisting unobtrusively within ecosystems. Yet, robustness alone is insufficient.

Natural dynamics are not merely obstacles, but potential resources for innovation. For example, could dew be harvested for cooling? Could daily temperature swings drive thermoelectric energy? Could plant movements generate cryptographic randomness? We believe NatureHCI researchers can start exploring these questions to understand how to leverage environmental variability, and open up new NatureHCI horizons. Among the most pressing challenges we identified was energy autonomy. Conventional batteries violated leave-no-trace ethics and require replacement, while solar cells depend on rare materials. We consider that NatureHCI researchers need to explore alternative energy pathways, such as piezoelectric rain capture, microbial fuel cells, and gravitational energy storage. Although individually modest, these methods hinted at hybrid low-power ecosystems for sustainable computing.

Finally, the problem of communication in connectivity-scarce environments was a recurrent topic. Without cellular or stable satellite links, we see the need to investigate delay-tolerant interactions and opportunistic networking, where data travels through humans, animals, or natural signals such as bioluminescence or plant electrophysiology. These unconventional approaches could prioritize integration and endurance over immediacy.

Overall, developing technology for NatureCI demands more than technical fixes; it requires rethinking the relationship between technology and the living world. Hence, NatureHCI researchers could start addressing this challenge by investigating a research agenda spanning material resilience, ecological energy systems, bio-hybrid computing, and distributed communication.

7.3 The Challenge of Evaluation and Assessment of Human-Nature Interactions

How do we know if NatureHCI interventions succeed? Traditional HCI evaluation methods [30], such as usability tests, performance metrics, and user satisfaction scores, fail to capture the complexity of human-nature relationships. Natural settings resist laboratory control. Effects may take years to manifest. Success might mean different things to humans, non-human organisms, and ecosystems. Hence, there is a lack of knowledge on how NatureHCI can develop evaluation approaches that are adequate to this complexity.

We recognize that standard HCI evaluation assumptions don't hold in natural settings. Controlled experiments require controlling variables, but nature resists control. Replication requires consistent conditions, but no two forests are identical. Statistical power requires large sample sizes, but deploying hundreds of devices might harm the ecosystems we are trying to help. Particularly, longitudinal evaluation presents special challenges. While a usability study might last hours, understanding NatureHCI's impacts might require years [9]. For example, does a technology that initially increases nature connection maintain that effect over time? Do systems deployed to support conservation actually improve ecosystem health? These questions cannot be answered through brief studies.

NatureHCI researchers could adapt methods from ecology and conservation biology, using before-after-control-impact designs that compare sites with and without interventions over extended periods [12]. Moreover, NatureHCI researchers have started to employ adaptive management approaches that treat deployment as an ongoing experiment, continuously adjusting based on observed outcomes. Another evaluation method is to explore participatory evaluation where communities define success metrics based on local values and monitor progress themselves [47].

Although multi-stakeholder evaluation frameworks attempt to address the benefit balance between human-nature relationships by assessing impacts across different groups [5], such as humans of various backgrounds, different species, and ecosystem processes; this multiplies complexity. For example, it is unclear how do we compare stress reduction in humans against stress increase in wildlife [56, 38], or how do we balance indigenous cultural values against scientific conservation goals?

The challenge of assessing human-nature relationships extends to developing new constructs and measures specific to NatureHCI. Existing scales for nature connection, developed primarily through correlational studies [14], may not capture how technology mediates the novel interactions NatureHCI designers are proposing. Hence, we need new instruments sensitive to the unique dynamics of technologically mediated nature experiences. Some researchers have explored novel evaluation approaches aligned with NatureHCI's values. Phenomenological methods attempt to capture the qualitative richness of nature experiences [57]. Multispecies ethnography follows effects across human and non-human participants [13]. Arts-based methods use creative expression to surface impacts that resist quantification [6]. These approaches sacrifice traditional notions of objectivity for deeper insight into lived experience.

In summary, to start addressing this challenge, NatureHCI researchers can start investigating how to develop reliable, nuanced methods to assess how technology impacts a person's subjective experience of spending time in nature, how to continue to evolve methods and tools to evaluate impacts on nature connections, and develop new methods, scales, and evidentiary standards for multidisciplinary evaluation of non-human impacts, drawing on environmental assessment, life-cycle assessment, and more-than-human methods.

7.4 The Challenge to Design Rich, Embodied and Multisensory Experiences in Nature

Nature experiences are inherently multisensory and embodied, yet digital representations of nature and mediated experiences are are frequently criticised for their sensory and affective impoverishment, lacking the richness and embodied depth of direct encounters [58]. This gap poses a major challenge for NatureHCI, which seeks not only to enhance sensorial and

embodied experiences through technology, but also to extend human perception beyond its natural limits. While such extensions offer possibilities for deeper understanding and connection with nonhuman worlds, they demand speculative and epistemic leaps to imagine, sense, and represent the life-worlds of nonhumans – what Jewitt et al. describe as "an entangled digital umwelt" [25].

A central difficulty lies in the visual bias of digital media. There is growing critical recognition that digital representations of the world privilege the visual [25, 40], but nature experiences are multisensory and embodied. Studies have begun to incorporate varied sensory experiences into virtual nature experiences, most commonly smell [63, 31, 52]; and touch [51]. This is important because multiple sensory pathways are (probably) involved in benefits of nature experiences [17], especially sound, smell, taste, touch. However, integrating these modalities in coherent, contextually responsive ways remains an open challenge.

More-than-human design frameworks push these concerns further by treating embodied encounters with animals, plants, weather, and land as active design materials rather than static contexts. For example, [3] highlight Indigenous concepts of relational accountability and embodied co-presence with land in participatory design. In this vein, work at the intersection of NatureHCI and OutdoorsHCI foregrounds the locative, embodied interactions of hikers, runners, etc. with their environment [27]. Rao et al. [45] call for reconceptualisation of smart building design around biophilic interactions, in ways that promote greater attention to embodied sensations and diverse sensory experiences. Still, the challenge persists: how can technologies genuinely attune us to more-than-human bodies and environments, rather than abstracting or instrumentalising them? Emerging work in HCI has started to surface hidden dimensions of embodiment by exploring experiences below the threshold of consciousness. For example, recent work has explored the possibility of surfacing human-gut interactions for reflection, well-being and "visceral conversations" [42], and circadian rhythms [2]. Notably, more research is needed into the multisensory pathways involved in beneficial nature experiences and [17]. For example, despite HCI attention to the experience of being in forests [41] and broader research into forest bathing , there is a lack of HCI research exploring interactions with phytoncides (antimicrobial volatiles emitted by plants) [17]. These omissions underscore a broader challenge: to design for sensory modalities and ecological processes that lie beyond conscious perception.

Jewitt [25] calls for a reimagining of digital communication technologies beyond the audiovisual paradigm, shaped by diverse cultural and ecological perspectives. This entails developing expanded, multisensory communication models, incorporating haptic, olfactory, and environmental sensing, to foster deeper forms of co-presence and more-than-human sociality. Through speculative and participatory design, NatureHCI can thus begin to map new sensory worlds – a "digital umwelt" that does not merely replicate human experience but reconfigures communication across species and ecologies. Within this broader agenda, sound emerges as a particularly promising yet underexplored frontier. As Bakker [7] illustrates, digital listening technologies are already transforming our ability to connect with elephants, sperm whales, bats, and bees, revealing how acoustic mediation might serve as a bridge toward richer, more reciprocal relationships with the more-than-human world.

Overall, NatureHCI researchers could start addressing these challenges by investigating how we can reimagine technologies for new ways of communicating and living with the more-than-human world and by creating methods and toolkits that support designers in engaging with the perceptual worlds of non-humans.

7.5 The Challenge of Leave-No-Trace Design

The technology industry's design ethos – rapid prototyping, constant iteration, and planned obsolescence – conflicts sharply with environmental ethics of minimizing impact [59]. Each discarded prototype embodies hidden costs: rare earth elements extracted through destructive mining, water-intensive data center operations, and electronic waste accumulating in ecosystems. For NatureHCI, this tension raises a central challenge: how can interactive systems be designed in alignment with leave-no-trace principles?

Addressing this requires rethinking core assumptions of the design process. Conventional HCI practices rely on disposability, assuming prototypes can be easily replaced or discarded. Yet in natural settings, there is no "away" – everything remains within ecological cycles. Researchers argued that design methods must evolve to prevent technological residues from persisting in the environment [26].

One approach involves prolonging low-fidelity prototyping. Rather than moving quickly to electronic models, designers could work longer with natural materials such as branches, grasses, clay, and sand – materials that can safely return to ecosystems when discarded. This extends design experimentation while reducing environmental footprint.

However, many NatureHCI systems depend on computational components that resist such substitution. This highlights the challenge of developing biodegradable electronics, including cellulose-based substrates, protein-derived polymers, and carbon conductors from renewable sources. Though still experimental, these innovations suggest pathways toward ecologically compatible computation [36].

Energy systems pose another critical difficulty. Conventional batteries and solar panels rely on toxic materials and complex manufacturing. NatureHCI researchers explored alternative, low-impact energy sources – wind, water flow, temperature differentials, microbial metabolism – each producing modest power but aligning with natural energy rhythms.

More radically, leave-no-trace might mean leaving no physical trace at all [60]. Could NatureHCI exist purely in software, using devices people already carry? Could interactive installations biodegrade or self-dismantle after use, ensuring full reintegration with their environment? Designing for disassembly and reuse further supports circular material lifecycles.

Finally, researchers emphasized that sustainability should not only minimize harm but also actively contribute to ecosystem health. Drawing on regenerative design, NatureHCI could envision systems that provide habitat, filter water, or enrich soil while sensing or computing. The goal thus shifts from leaving no trace to leaving beneficial traces – where technology becomes a participant in ecological regeneration rather than a source of waste.

7.6 The Challenge of Cultural Perspectives

NatureHCI risks perpetuating colonial patterns if it assumes universal relationships between humans, nature, and technology. Different cultures hold radically different concepts of what nature is, how humans should relate to it, and what role technology should play [29, 61]. How does NatureHCI respect and support this diversity while building coherent knowledge?

The challenge begins with recognizing that "nature" itself is a cultural construct. Many indigenous languages have no word separating nature from culture – the divide that seems fundamental to Western thought simply doesn't exist [46]. Some cultures see nature as kin requiring reciprocal relationships. Others view it as a resource for human use. Still others understand nature as teacher, adversary, or manifestation of the sacred [29, 20]. These different understandings lead to different technological approaches. A culture viewing nature

as kin might develop technologies facilitating communication and care. One seeing nature as teacher might create technologies for observation and learning. These aren't just different applications of neutral technology but fundamentally different technological paradigms.

Consider how different cultures approach the question of whether technology belongs in natural settings. Some embrace any tool that deepens understanding or connection. Others maintain strict boundaries between technological and natural spaces. Some integrate new technologies into traditional practices. Others see technology as inherently alienating from natural relationships. NatureHCI cannot assume any single approach is correct [21].

This challenge requires developing what we might call "ontological flexibility" – the ability to work within different worldviews without imposing our own. This goes beyond surface-level cultural sensitivity to engaging with different fundamental assumptions about reality. It means recognizing that other knowledge systems aren't primitive versions of Western science but sophisticated ways of understanding that may grasp aspects of nature that Western approaches miss [29].

Indigenous knowledge systems offer particular insight for NatureHCI [1]. Many indigenous peoples have developed sophisticated technologies for living sustainably within specific ecosystems over millennia [10]. These technologies – from fish weirs that selectively harvest while maintaining populations to controlled burns that enhance ecosystem health – demonstrate deep integration between human needs and natural cycles. They suggest possibilities for NatureHCI systems that work with rather than against natural processes.

But engaging with indigenous knowledge raises critical questions about intellectual property, benefit sharing, and cultural appropriation [11]. Too often, Western researchers have extracted traditional knowledge without reciprocating benefits or respecting cultural protocols. NatureHCI must develop ethical frameworks ensuring that collaboration genuinely benefits indigenous communities rather than exploiting their knowledge.

The challenge extends to supporting multiple cultural approaches within single systems. A NatureHCI platform might need to accommodate users who want detailed scientific data alongside those seeking spiritual connection, those focused on resource extraction alongside those practicing reciprocal care. Rather than averaging these differences into bland universality, how do we create systems that support multiple valid approaches?

7.7 The Challenge of Accessibility and Inclusion in Nature

Accessibility and inclusion are essential to ensure that everyone – regardless of physical, cultural, or socio-economic background – can engage with nature through technology. NatureHCI offers a unique opportunity to augment, mediate, and transform nature experiences for people who face systemic barriers to access, including physical impairments, mobility constraints, exclusion of Indigenous knowledge systems, and inequitable resource distribution [22, 55].

However, designing inclusive nature-based technologies is far from straightforward. NatureHCI's mission, to explore and shape human–technology–nature relationships, positions it to address such barriers, yet research in this area often assumes normative bodies, capabilities, and contexts, inadvertently marginalizing those whose connection to and stewardship of nature are most vital [54]. To ensure equitable participation, NatureHCI must consider constraints such as cost, time, mobility, policy, and uneven technological access. Without deliberate attention, the field risks reproducing the exclusionary patterns observed in other HCI domains.

The challenge is systemic. Barriers to accessing both nature and technology are rooted in structural inequities, including policies, infrastructure, and socio-economic disparities that cannot be solved by isolated technological interventions [53]. For example, even individuals without impairments face economic and geographic barriers to aquatic recreation [24], where equipment and travel costs restrict access to coastal or freshwater environments. Meanwhile, aquatic ecosystems are often home to Indigenous and Aboriginal communities for whom tourism-based or recreational engagement may represent a cultural and environmental trade-off [48]. Thus, accessibility in NatureHCI cannot rely on a universal model; it must account for diverse and context-specific needs.

In the broader HCI community, research on disability and accessibility has advanced significantly in recent years, particularly around ageing, mobility, and visual impairments, yet NatureHCI research remains fragmented, often limited to specific sites, abilities, or technologies. Examples include voice descriptions for nature imagery to aid visually impaired users, or Alsaleem et al. (2020)'s work on augmented navigation for paraplegic skiers [4]. Studies addressing social interaction phobias [43] and virtual nature experiences [28] demonstrate potential to expand accessibility for those with mobility or socio-economic constraints, though such approaches are still emerging.

To move forward, NatureHCI researchers could start investigating how to balance nature access and environmental protection by partnering with local communities. For example, in Colombia, a conservation area alternates open-access tourism periods with closures for ecological recovery [62]. Researchers can also examine how intersecting barriers, disability, mobility, economic, and policy constraints, jointly shape accessibility to both technology and nature, and how to engage with people experiencing accessibility barriers to co-design meaningful, self-directed engagements with nature.

In summary, addressing accessibility and inclusion in NatureHCI requires rethinking beyond individual devices or interfaces toward reshaping the systemic conditions that determine who can access and benefit from nature–technology interactions. Through collaboration with Indigenous leaders, policymakers, and local communities, NatureHCI can create frameworks that open both nature and technology to those historically excluded from them.

7.8 Synthesis: Challenges as Invitations

These grand challenges interconnect in ways that prevent addressing them in isolation. Designing for non-human stakeholders requires cultural sensitivity to different ways of understanding non-human agency. Leave-no-trace design must account for unpredictability that might damage delicate prototypes. Evaluation methods must assess impacts across stakeholders while respecting cultural differences in what constitutes benefit or harm.

Rather than viewing this interconnection as complication, we might see it as invitation to develop truly integrated approaches. The challenges call for new forms of collaboration crossing disciplinary and cultural boundaries. They demand methods that work with rather than against complexity. They require patience, humility, and acceptance of irreducible uncertainty.

Most fundamentally, these challenges invite us to reconsider HCI's basic assumptions when applied to nature contexts. The field's roots in making technology useful and usable for humans may need fundamental expansion to include making technology beneficial for the more-than-human world. This doesn't mean abandoning human-centered values but situating them within larger ecological contexts.

The challenges also reveal NatureHCI's potential contributions beyond specific applications. By grappling with non-human stakeholders, the field pushes HCI to consider more diverse forms of agency and intelligence. By addressing temporal mismatches, it explores how technology might support rather than accelerate human relationships with time. Engaging cultural diversity demonstrates possibilities for pluralistic rather than universal design.

Perhaps most importantly, these challenges position NatureHCI as a field of hope in the anthropocene [44]. Rather than depicting technology and nature as inevitably opposed, they outline possibilities for beneficial integration. Rather than lamenting what's been lost, they focus on what might be cultivated. Rather than prescribing single solutions, they embrace the diversity of approaches needed for diverse contexts.

As participants repeatedly noted during the seminar, these truly are "grand" challenges – they will not be solved by single researchers or even single institutions. They require sustained collective effort over years and decades. But their grandness also makes them worthy of our best efforts. In addressing them, we address some of the most pressing questions of our time: How can humanity live sustainably on Earth? How can technology serve rather than subvert ecological wellbeing? How can diverse cultures collaborate while maintaining their distinctiveness? The challenges thus serve not as discouragements but as beacons, guiding NatureHCI toward futures where human and natural flourishing support rather than oppose each other.

8 Research Agenda and Future Directions

The seminar's final sessions transformed insights and aspirations into concrete plans. Participants recognized that without specific commitments, even the most inspiring discussions would dissipate upon returning to daily academic pressures. The research agenda that emerged balanced an ambitious vision with pragmatic steps.

8.1 Immediate Actions and Short-term Priorities

Before leaving Dagstuhl, participants made specific commitments with deadlines and designated leads. The immediate priority was submitting a comprehensive grand challenges paper to CHI 2026. Unlike typical conference papers authored by small teams, this would represent the full community's collective intelligence. Writing groups formed around each major challenge, with coordinators ensuring coherence across sections. Monthly video calls would maintain momentum, with drafts circulating for feedback. The paper would acknowledge all seminar participants as contributors while identifying core authors who led writing efforts.

Establishing NatureHCI's presence at major conferences emerged as another immediate priority. Beyond the grand challenges paper, participants committed to proposing workshops at CHI, DIS, TEI, and Ubicomp. Each workshop would focus on different aspects, such as methods, ethics, technology, and applications, while building toward a coherent research program. Workshop proposals would be coordinated to avoid overlap and ensure progression across venues.

The methods repository is aimed to be launched by December 2025. Rather than waiting for perfect infrastructure, they would begin with a simple shared platform, adding sophistication as the collection grew. Each method would include theoretical grounding, practical instructions, case studies of use, and reflections on limitations. Review processes would ensure quality while encouraging experimentation.

The Augmented Animals Special Interest Group emerged unexpectedly from a final-day comment that captured imaginations. When organizer Inami suggested that "augmented humans are now too small, maybe we can say augmented animals," participants immediately began envisioning possibilities. Unlike other groups that formed early and met throughout the week, this group crystallized in the final hours, demonstrating the seminar's generative environment. Despite forming late, the group generated remarkable momentum. They planned a workshop at the Augmented Humans conference to explore the concept further. They identified potential collaborators in veterinary science, wildlife biology, and animal cognition. They drafted ethical guidelines specific to animal augmentation. Most importantly, they demonstrated how NatureHCI could radically expand conceptions of who benefits from interactive technology.

8.2 Medium-term Development

Over the next two to three years, participants envisioned the gradual institutionalization of NatureHCI within the HCI community. This development may begin through special tracks or workshops embedded within established conferences, with the longer-term goal of forming a dedicated symposium. In parallel, academic programs could introduce NatureHCI content into curricula, initially through special-topics seminars and subsequently through certificate programs or specialized degree pathways.

Research infrastructure is expected to expand from individual projects toward shared and sustained resources. Field stations designed for NatureHCI research could provide controlled environments for longitudinal studies. Distributed sensor networks would enable cross-site data comparison, while fabrication facilities focused on sustainable prototyping and materials innovation would support experimental development. Although these initiatives would require substantial investment, they would establish the foundation for systematic, interdisciplinary research.

To sustain this growth, participants identified the need for diversified funding strategies. Large-scale collaborative grants could support infrastructure establishment and shared datasets, while smaller grants may foster exploratory or proof-of-concept projects. Partnerships with industry may provide technical or material resources, contingent on maintaining research independence. Foundation and philanthropic support could address equity, conservation, and community engagement goals, while government agencies may recognize NatureHCI's relevance to climate resilience, environmental stewardship, and public health.

8.3 Long-term Vision

Looking toward 2035, participants envisioned NatureHCI evolving from an emerging topic into an established research field with measurable social and ecological impact. Its principles would be embedded in mainstream technology design, where ecosystem considerations accompany user experience. Conservation organizations would routinely employ NatureHCI methods, and universities would graduate researchers fluent in both computational and ecological thinking. At a broader level, NatureHCI was imagined to contribute to transformative shifts in human–nature relations. Technologies designed with its principles would enable urban populations to sustain connections with nature, while supporting rural and Indigenous communities in sharing local knowledge and maintaining cultural integrity. Global sensor networks could advance ecosystem understanding while upholding data sovereignty and ethical

governance. By this stage, the field would have developed robust theoretical, methodological, and ethical frameworks for studying and designing technology in natural contexts. Crucially, NatureHCI would exemplify alternatives to extractive technology paradigms, fostering reciprocity rather than exploitation, sustainability rather than consumption, and connection rather than separation. Through such approaches, the field would demonstrate how technology can participate in regenerative rather than destructive relationships with the natural world.

8.4 A Living Agenda

Participants emphasized that the NatureHCI research agenda should remain a living framework rather than a fixed plan. Regular community gatherings would reassess priorities in light of emerging insights and changing environmental contexts. Inclusivity was seen as essential – actively welcoming new voices to sustain diversity and prevent institutional stagnation. Equally important was a commitment to transparency: acknowledging and learning from failures while sharing successful practices openly across the community. This adaptive orientation drew inspiration from ecological principles. Like resilient ecosystems, the research community would thrive through diversity, flexibility, and collaboration. Such a structure would enable NatureHCI to evolve organically and sustain long-term inquiry across disciplines and geographies.

Ultimately, participants described NatureHCI not merely as another research domain, but as a collective response to broader planetary challenges. The agenda embodies a commitment to engaged, responsible scholarship that bridges technology, ecology, and society – establishing both a shared sense of purpose and a practical foundation for continued development.

9 Conclusion: Seeds Planted at Dagstuhl

As participants prepared to depart Schloss Dagstuhl on the final morning, a palpable sense of accomplishment mixed with anticipation filled the air. The week had exceeded expectations, transforming thirty individuals with overlapping interests into a coherent community with shared purpose. Yet everyone recognized this was beginning rather than culmination – seeds planted that would require careful tending to flourish. Finally, the contributions of this seminar are: Intellectual Contribution: The seminar defined NatureHCI as the study of how interactive computing mediates human–nature engagement, distinguishing it from both OutdoorHCI and SportsHCI. A grand challenges framework outlined enduring research tensions, such as designing for nonhuman stakeholders and aligning prototyping with leave-no-trace ethics. Methodological innovations like phenology circles and multispecies personas demonstrated that new epistemic approaches are required, not mere adaptations. Cross-disciplinary integration enriched understanding, linking computing, design, ecology, and Indigenous knowledge. Community Formation: The event transformed shared interest into an emerging research community. Working groups coalesced into functional teams, strengthened by global diversity in geography, discipline, and career stage. Participants emphasized maintaining openness and inclusivity by recruiting underrepresented voices, ensuring the field remains plural rather than exclusive. Practical Outcomes: Concrete outputs ensured momentum: a CHI 2026 grand challenges paper, a methods repository, and pilot studies to test ideas across contexts. These pragmatic activities, such as publishing, teaching, funding, were recognized as essential for institutionalizing NatureHCI. Challenges Acknowledged:

Participants noted systemic obstacles: academic incentives favoring individual over collective work, disciplinary funding silos, and the urgency of environmental change. Ongoing tensions, innovation vs. environmental critique, universality vs. locality, were reframed as productive forces sustaining intellectual vitality. Vigilance against commercial co-optation was deemed crucial to preserve the field's ethical grounding.

Ultimately, the seminar's legacy lies in cultivating a field that unites technological innovation with ecological care, laying the foundation for sustained, responsible research into human–technology-nature relations.

10 Acknowledgements

We thank Dagstuhl for their extensive support and all the participants who contributed to this report as part of a collective effort. A particular thank you to Maria Fernanda Montoya for volunteering first to help and supporting the organizers in finalizing the report.

References

1 José Abdelnour-Nocera, Torkil Clemmensen, and Masaaki Kurosu. Reframing HCI Through Local and Indigenous Perspectives. *International Journal of Human–Computer Interaction*, 29(4):201–204, March 2013.

2 Saeed Abdullah, Mark Matthews, Elizabeth L Murnane, Geri Gay, and Tanzeem Choudhury. Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 673–684, 2014.

3 Yoko Akama, Penny Hagen, and Desna Whaanga-Schollum. Problematizing Replicable Design to Practice Respectful, Reciprocal, and Relational Co-designing with Indigenous People. *Design and Culture*, 11(1):59–84, January 2019.

4 Ahmad Alsaleem, Ross Imburgia, Andrew Merryweather, Jeffery Rosenbluth, Stephen Trapp, and Jason Wiese. *Creating a User-Controllable Skiing Experience for Individuals with Tetraplegia*, pages 275–290. Springer International Publishing, Cham, 2020.

5 Angelina Aspra Aquino, Ian Mongunu Gumbula, Nicola Bidwell, and Steven Bird. What's the weather story? Both-ways learning in Indigenous-led climate communication workshops in northern Australia. In *Proceedings of the Participatory Design Conference 2024: Exploratory Papers and Workshops - Volume 2*, volume 2 of *PDC '24*, pages 166–174, New York, NY, USA, August 2024. Association for Computing Machinery.

6 Bronwyn Bailey-Charteris. *The Hydrocene: Eco-aesthetics in the age of water*. Taylor & Francis, 2024.

7 Saskia Bakker and Karin Niemantsverdriet. The interaction-attention continuum: Considering various levels of human attention in interaction design. *International Journal of Design*, 10(2):1–14, 2016.

8 James Barber, Yiting Wen, Aaron Ye, and Austin Allen. A nature hci approach to intergenerational icebreaking. In *Proceedings of the 2025 Conference on Creativity and Cognition*, pages 498–502, 2025.

9 Fabien Barthelot, Marc Le Goc, and Eric Pascual. Influence of seasons on human behavior in smart environments. In *International Workshop on Ambient Assisted Living*, pages 146–151. Springer, 2015.

10 N. J. Bidwell and H. Winschiers-Theophilus. What Indigenous Knowledge is Not: An Introductory Note. *At the Intersection of Indigenous and Traditional Knowledge and Technology Design*, page 15, 2015.

11 Nicola J Bidwell, Peta-Marie Standley, Tommy George, and Vicus Steffensen. The landscape's apprentice: lessons for place-centred design from grounding documentary. In *Proceedings of the 7th ACM conference on Designing interactive systems*, DIS '08, pages 88–98, New York, NY, USA, February 2008. Association for Computing Machinery.

12 Eli Blevis, Susanne Bødker, John Flach, Jodi Forlizzi, Heekyoung Jung, Victor Kaptelinin, Bonnie Nardi, and Antonio Rizzo. Ecological perspectives in hci: Promise, problems, and potential. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2401–2404, 2015.

13 Ashley Boone, Christopher A. Le Dantec, and Carl DiSalvo. Embodied traces: Multispecies entanglement in urban spaces. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 1074–1086, New York, NY, USA, 2024. Association for Computing Machinery.

14 Elodie Bouzekri and Guillaume Rivière. Choosing a questionnaire measuring connectedness to nature for human-computer interaction user studies. In *Actes de la 33e conférence internationale francophone sur l'Interaction Humain-Machine (IHM'22)*. ACM, 2022.

15 Michelle Chang, Chenyi Shen, Aditi Maheshwari, Andreea Danielescu, and Lining Yao. Patterns and opportunities for the design of human-plant interaction. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 925–948, 2022.

16 Keith Cheverst, Mads Bødker, and Florian Daiber. Technology and mastery: exploring design sensitivities for technology in mountaineering. In *HCI Outdoors: Theory, Design, Methods and Applications*, pages 197–211. Springer, 2020.

17 Lara S. Franco, Danielle F. Shanahan, and Richard A. Fuller. A review of the benefits of nature experiences: More than meets the eye. *International Journal of Environmental Research and Public Health*, 14(8), 2017.

18 Jonna Häkkilä, Keith Cheverst, Johannes Schöning, Nicola J Bidwell, Simon Robinson, and Ashley Colley. Naturechi: unobtrusive user experiences with technology in nature. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3574–3580, 2016.

19 Mahmoud Hassan, Florian Daiber, Frederik Wiehr, Felix Kosmalla, and Antonio Krüger. Footstriker: An ems-based foot strike assistant for running. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(1):1–18, 2017.

20 Hanna Helander, Henna Aikio, Pigga Keskitalo, and Tuija Turunen. Land-based participatory pedagogical experiment in sami language distance teaching: Maintaining children's relationships with land and nature. In *Handbook of research on teaching in multicultural and multilingual contexts*, pages 1–22. IGI Global Scientific Publishing, 2022.

21 Linda Hirsch, Siiri Paananen, Denise Lengyel, Jonna Häkkilä, Georgios Toubekis, Reem Talhouk, and Luke Hespanhol. Human–computer interaction (hci) advances to re-contextualize cultural heritage toward multiperspectivity, inclusion, and sensemaking. *Applied Sciences*, 14(17), 2024.

22 Fay Holland. Out of Bounds. Technical report, Groundwork, 2021.

23 Kuisma Hurtig, Jemina Colley, Michael Jones, and Jonna Häkkilä. Glove navigator for skiing in the mountains. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, pages 509–511, 2023.

24 Gail H Ito. Barriers to swimming and water safety education for african americans. *International Journal of Aquatic Research and Education*, 8(3):4, 2014.

25 Carey Jewitt. Provocation: stop replicating and start reimagining digital communication for the sensing body. *The Senses and Society*, pages 1–13, 2025.

26 Chen Jin, Luyi Yang, and Cungen Zhu. Right to repair: Pricing, welfare, and environmental implications. *Manage. Sci.*, 69(2):1017–1036, February 2023.

27 Michael D Jones, Zann Anderson, Jonna Häkkilä, Keith Cheverst, and Florian Daiber. Hci outdoors: understanding human-computer interaction in outdoor recreation. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2018.

28 Tuomas Kari, Ann Ojala, Mika Kurkilahti, and Liisa Tyrväinen. Comparison between three different delivery technologies of virtual nature on psychological state related to general stress recovery: An experimental study. *Journal of Environmental Psychology*, 100:102452, 2024.

29 Robin Wall Kimmerer. *Braiding Sweetgrass: Indigenous Wisdom, Scientific Knowledge and the Teachings of Plants*. Milkweed Editions, New York, 2013.

30 Gitte Lindgaard. The usefulness of traditional usability evaluation methods. *Interactions*, 21(6):80–82, 2014.

31 Marilia K. S. Lopes and Tiago H. Falk. Audio-visual-olfactory immersive digital nature exposure for stress and anxiety reduction: a systematic review on systems, outcomes, and challenges. *Frontiers in Virtual Reality*, 5, 2024.

32 Hong Luo, Tuomas Kari, Rakesh Patibanda, Maria Fernanda Montoya, Josh Andres, Don Samitha Elvitigala, and Florian 'Floyd' Mueller. Plantmate: A bidirectional touch-based system for enhancing human-plant empathy and pro-environmental behavior. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

33 Maria F Montoya, Aryan Saini, Sarah Jane Pell, Phoebe O Toups Dugas, and Florian 'Floyd' Mueller. Exploring the role of interactive technology to enrich surfing. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 3581–3599, 2025.

34 Natasha Myers. Conversations on plant sensing: Notes from the field. *Nature*, 3:35–66, 2015.

35 Scott Newey, Paul Davidson, Sajid Nazir, Gorry Fairhurst, Fabio Verdicchio, R Justin Irvine, and René Van Der Wal. Limitations of recreational camera traps for wildlife management and conservation research: A practitioner's perspective. *Ambio*, 44(Suppl 4):624–635, 2015.

36 Madalina Nicolae, Vivien Roussel, Marion Koelle, Samuel Huron, Jürgen Steimle, and Marc Teyssier. Biohybrid devices: Prototyping interactive devices with growable materials. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, pages Article 31, 1–15. Association for Computing Machinery, 2023.

37 Leticia Antunes Nogueira, Maiken Bjørkan, and Brigt Dale. Conducting research in a post-normal paradigm: Practical guidance for applying co-production of knowledge. *Frontiers in Environmental Science*, 9:699397, 2021.

38 Emeline Nogues, Ava Arends, and Marina A. G. von Keyserlingk. Rapid systematic literature review: Camera trap sampling in ecological studies: Considerations of wildlife welfare. *Animal Welfare*, 34:e44, January 2025.

39 Ralf Oelmüller. Interplant communication via hyphal networks. *Plant Physiology Reports*, 24(4):463–473, 2019.

40 David. Parisi. *Archaeologies of Touch: Interfacing with Haptics from Electricity to Computing*. University of Minnesota Press, 2018.

41 Bum Jin Park, Yuko Tsunetsugu, Tamami Kasetani, Takahide Kagawa, and Yoshifumi Miyazaki. The physiological effects of shinrin-yoku (taking in the forest atmosphere or forest bathing): evidence from field experiments in 24 forests across japan. *Environmental health and preventive medicine*, 15(1):18–26, 2010.

42 Nandini Pasumarthy, Shreyas Nisal, Jessica Danaher, Elise van den Hoven, and Rohit Ashok Khot. Go-go biome: Evaluation of a casual game for gut health engagement and reflection. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

43 Maaret Posti, Johannes Schöning, and Jonna Häkkilä. Unexpected journeys with the hobbit: the design and evaluation of an asocial hiking app. In *Proceedings of the 2014 conference on Designing interactive systems*, pages 637–646, 2014.

44 Alison Pritchard and Miles Richardson. *The Relationship Between Nature Connectedness and Human and Planetary Wellbeing: Implications for Promoting Wellbeing, Tackling Anthropogenic Climate Change and Overcoming Biodiversity Loss*, pages 71–84. Springer International Publishing, Cham, 2022.

45 Shruti Rao, Judith Good, and Hamed Alavi. Lessons from biophilic design: Rethinking affective interaction design in built environments. 2025.

46 Graeme Reed, Nicolas D Brunet, Deborah McGregor, Curtis Scurr, Tonio Sadik, Jamie Lavigne, and Sheri Longboat. There is no word for 'nature'in our language: rethinking nature-based solutions from the perspective of indigenous peoples located in canada. *Climatic Change*, 177(2):32, 2024.

47 Katherine Reilly, Gillian Russell, Lauren Thu, and Jihyun Park. Situated design and false creek futures: Relationality engagement and creativity in eco-social information systems. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, page 944–958, New York, NY, USA, 2025. Association for Computing Machinery.

48 Malwina Schafft, Benjamin Wegner, Nora Meyer, Christian Wolter, and Robert Arlinghaus. Ecological impacts of water-based recreational activities on freshwater ecosystems: a global meta-analysis. *Proceedings of the Royal Society B*, 288(1959):20211623, 2021.

49 DJ Shepherdson, KC Carlstead, and N Wielebnowski. Cross-institutional assessment of stress responses in zoo animals using longitudinal monitoring of faecal corticoids and behaviour. *Animal Welfare*, 13(S1):S105–S113, 2004.

50 Alessandro Soro, Margot Brereton, Tshering Dema, Jessica L Oliver, Min Zhen Chai, and Aloha May Hufana Ambe. The ambient birdhouse: An iot device to discover birds and engage with nature. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

51 Pia Spangenberger, Sarah-Christin Freytag, and Sonja M. Geiger. Embodying nature in immersive virtual reality: Are multisensory stimuli vital to affect nature connectedness and pro-environmental behaviour? *Computers & Education*, 212:104964, 2024.

52 Pia Spangenberger, Kilian Sanchez-Holguin, and Sarah-Christin Freytag. Scent box: Prototyping and instructions for olfactory enhancement of vr-experiences. In Jule M. Krüger, Daniela Pedrosa, Dennis Beck, Marie-Luce Bourguet, Andreas Dengel, Rami Ghannam, Alan Miller, Anasol Peña-Rios, and Jonathon Richter, editors, *Immersive Learning Research Network*, pages 18–33, Cham, 2025. Springer Nature Switzerland.

53 Katta Spiel, Kathrin Gerling, Cynthia L. Bennett, Emeline Brulé, Rua M. Williams, Jennifer Rode, and Jennifer Mankoff. Nothing About Us Without Us: Investigating the Role of Critical Disability Studies in HCI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA, April 2020. Association for Computing Machinery.

54 Cella M Sum, Rahaf Alharbi, Franchesca Spektor, Cynthia L Bennett, Christina N Harrington, Katta Spiel, and Rua Mae Williams. Dreaming Disability Justice in HCI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, pages 1–5, New York, NY, USA, April 2022. Association for Computing Machinery.

55 Yan Sun, Somidh Saha, Heike Tost, Xiangqi Kong, and Chengyang Xu. Literature Review Reveals a Global Access Inequity to Urban Green Spaces. *Sustainability*, 14(3):1062, January 2022.

56 Don E. Swann, Kae Kawanishi, and Jonathan Palmer. Evaluating Types and Features of Camera Traps in Ecological Studies: A Guide for Researchers. In Allan F. O'Connell,

James D. Nichols, and K. Ullas Karanth, editors, *Camera Traps in Animal Ecology: Methods and Analyses*, pages 27–43. Springer Japan, Tokyo, 2011.

57 Oscar Tomico, Anton Poikolainen Rosén, Svenja Keune, Ferran Altarriba Bertran, Danielle Wilde, Daniel Fernández Galeote, Tau Ulv Lenskjold, Ruut Tikkanen, Oğuz 'Oz Buruk, and Velvet Spors. Seeding a repository of methods-to-be for nature-entangled design research. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 1101–1115, New York, NY, USA, 2024. Association for Computing Machinery.

58 Minh-Xuan A Truong and Susan Clayton. Technologically transformed experiences of nature: A challenge for environmental conservation? *Biological Conservation*, 244:108532, 2020.

59 Eldy S Lazaro Vasquez, Hao-Chuan Wang, and Katia Vega. The Environmental Impact of Physical Prototyping: a Five-Year CHI Review. 2020.

60 Julia Watson. *Lo-TEK. Design by Radical Indigenism.* Taschen GmbH, Cologne Paris, 2019.

61 Sarah Webber, Ryan M Kelly, Greg Wadley, and Wally Smith. Engaging with nature through technology: A scoping review of hci research. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–18, 2023.

62 Bradley Wilson, Juan C Londono, Jovelyn Ferrer, and Bastian Popp. Colombia's tayrona national park: recommendations for future regional development. In *Handbook on Tourism and Behaviour Change*, pages 250–268. Edward Elgar Publishing, 2023.

63 Muhammed Yildirim, Anastasia Globa, Ozgur Gocer, and Arianna Brambilla. Digital smell technologies for the built environment: Evaluating human responses to multisensory stimuli in immersive virtual reality. *Building and Environment*, 271:112608, 2025.

Participants

- Ahmad Alsaleem
University of Utah – Salt Lake City, US
- Natalie Andrus
Virginia Polytechnic Institute – Blacksburg, US
- Bill Borrie
Deakin University – Melbourne, AU
- Margot Brereton
Queensland University of Technology – Brisbane, AU
- Florian Daiber
DFKI – Saarbrücken, DE
- Don Samitha Elvitigala
Monash University – Melbourne, AU
- Masahiko Inami
University of Tokyo, JP
- Carey Jewitt
University College London, GB
- Michael Jones
Brigham Young University – Provo, US
- Tuomas Kari
National Resources Institute Finland – Helsinki, FI
- Hong Luo
Monash University – Melbourne, AU
- Andrii Matviienko
KTH Royal Institute of Technology – Stockholm, SE
- Scott McCrickard
Virginia Tech – Blacksburg, US
- Maria Fernanda Montoya Vega
Monash University – Melbourne, AU
- Florian 'Floyd' Mueller
Monash University – Melbourne, AU
- Siiri Paananen
University of Lapland – Rovaniemi, FI
- Nandini Pasumarthy
Monash University – Melbourne, AU
- Rakesh Patibanda
Monash University – Melbourne, AU
- Ambika Shahu
IT:U Interdisciplinary Transformation University – Linz, AT
- Sarah Webber
The University of Melbourne, AU
- Jason Wiese
University of Utah – Salt Lake City, US

Report from Dagstuhl Seminar 25311

Generative AI in Programming Education

Michelle Craig*[1], Paul Denny*[2], Natalie Kiesler*[3], and James Prather*[4]

1 **University of Toronto, CA.** mcraig@cs.toronto.edu
2 **University of Auckland, NZ.** p.denny@auckland.ac.nz
3 **Technische Hochschule Nürnberg, DE.** natalie.kiesler@th-nuernberg.de
4 **Abilene Christian University, US.** james.prather@acu.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25311 "Generative AI in Programming Education". During the seminar, we examined the transformative impact of Generative AI on programming education. Because they can solve many introductory tasks given only natural language prompts, AI tools are challenging established approaches to programming education, in which there has been a traditional emphasis on writing small programs and providing (automated) feedback to learners. While these developments raise concerns about student over-reliance and inaccurate feedback, they also open opportunities for new pedagogical practices, such as fostering prompt literacy, adapting curricula, and designing AI-assisted learning tools. The present seminar convened 42 international experts to exchange knowledge, present research, and share innovations through keynotes, lightning talks, and tool demonstrations. Collaborative working groups explored implications for learning outcomes, assessment, equity, human values, and tool design, while identifying directions for systematic evaluation and interdisciplinary research. The seminar successfully established a foundation for a sustained community of practice and set an agenda for advancing programming education in the era of Generative AI.

Seminar July 27 – August 1, 2025 – https://www.dagstuhl.de/25311
2012 ACM Subject Classification Human-centered computing; Social and professional topics → Computing education
Keywords and phrases artificial intelligence, computer programming, computing education, generative ai, large language models
Digital Object Identifier 10.4230/DagRep.15.7.253

1 Executive Summary

Natalie Kiesler (Technische Hochschule Nürnberg, DE)
Michelle Craig (University of Toronto, CA)
Paul Denny (University of Auckland, NZ)
James Prather (Abilene Christian University, US)

Generative AI stands to significantly disrupt education in general and programming education is no exception. In addition, learning to program has several unique requirements and characteristics that require specific approaches. Evidence from the past several decades on how humans learn programming supports the commonly adopted approach of having students write many small programs. Often these are checked, and feedback is provided, by automated assessment tools. However, Generative AI has likely rendered this approach

* Editor / Organizer

obsolete given that easy-to-use tools are now readily available that can solve introductory computing problems with natural language prompts. At the same time, it is well known that the large language models that power Generative AI tools sometimes provide outputs that are either incorrect or inappropriate for the current understanding of a learner, raising concerns around student over-reliance and poor learning outcomes.

Educators are currently taking a variety of approaches, including ignoring the issue. Generative AI is a nascent, yet very rapidly developing field and new challenges and opportunities arise frequently making it extremely difficult for educators to keep pace with developments. Prototype tools that leverage Generative AI to facilitate learning are appearing, however most have yet to be deployed or adopted at scale. New pedagogical approaches are also emerging to foster the development of new kinds of skills, such as effective prompt creation, and new learning resources such as textbooks are appearing that teach programming hand in hand with Generative AI. Such approaches, however, have not yet been evaluated at scale as the field is developing so rapidly.

This Dagstuhl Seminar had the goal to bring together experts and stakeholders in Generative AI and computing education to foster collaboration and to chart a way forward as Generative AI continues to improve and proliferate. It was the goal of this seminar to leverage the experience and knowledge of dozens of programming education experts from around the world to form an enduring community of practice. During the seminar, we started to develop further strategies for incorporating LLMs into programming education and to rigorously evaluate their use and impact. We also explored the following topics in the context of Generative AI in programming education: accessibility; diversity, equality, and inclusion; resources; introductory programming for computer science majors and non-majors; advanced courses (that use programming); curriculum changes; novel pedagogies, approaches and tools; and industry use and changes that may lead to new learning outcomes.

These discussions were informed by participants' prior research and addressed the following objectives:

- Identify current implications of Generative AI on programming education, learning objectives, and curricula.
- Develop recommendations for the pedagogical integration of Generative AI in programming courses.
- Identify and establish interdisciplinary research objectives and questions to investigate Generative AI in programming education.

The seminar was structured into several sections, with all 42 participants actively involved: lightning and keynote talks followed by group brainstorming sessions, and dedicated workshop group sessions. At the beginning of the seminar, every attendee introduced themselves, and the seminar leaders provided an overview of the current GenAI landscape by introducing recent studies, emerging themes, trends, and tools. In addition, we had keynote presentations and discussions on AI in the curriculum, and specifically, Google's AI curriculum was presented. The keynote talks were delivered, as follows:

- Paul Denny: Opening Remarks: Generative AI and the Future of Programming Education
- Natalie Kiesler and James Prather: Beyond the Hype: A Sneak Peek into the Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools
- Dennis Bouvier: The Rest of the Robots: GenAI in Upper-level Computing
- Leo Porter and Daniel Zingaro: Learn AI-Assisted Python Programming
- James Prather: A New Curriculum for Computer Science that integrated GenAI: A Collaboration between Google and Academia
- Titus Winters: Generative AI in Software Engineering

We dedicated two more sessions to lightning talks and tool presentations. Prior to the seminar, every attendee had been invited to present recent findings, insights, or tools. This led to 20 lightning talks and tool presentations, which are represented in this report.

The lightning talk and tool sessions were accompanied by brainstorming sessions to identify recent challenges, opportunities, and future directions for research and practice. The brainstorming sessions also helped to form working groups. These subgroups focused on the following aspects: learning outcomes, assessment, social learning, CS1, meta-cognition, access and equity, quality use of GenAI, human values, introductory computing for non-majors, and AI-integrated learning tools. After two check-ins with the subgroups, several of them had merged, so that ultimately, three large working groups remained. The results of their discussions are summarized in this report.

2 Table of Contents

3 Overview of Talks

3.1 Feedback Quality Overview (open and state-of-the-art LLMs)

Imen Azaiz (LMU München, DE)

Joint work of Imen Azaiz, Natalie Kiesler, Sven Strickroth
Main reference Imen Azaiz, Natalie Kiesler, Sven Strickroth: "Feedback-Generation for Programming Exercises With GPT-4", in Proc. of the 2024 on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2024, Milan, Italy, July 8-10, 2024, ACM, 2024.
URL https://doi.org//10.1145/3649217.3653594

It's well known that in introductory computer science courses, not only do many students struggle with learning programming, but educators also often lack the resources to support them effectively, especially when it comes to providing timely, personalized feedback at scale. Large Language Models (LLMs) have shown potential to address this challenge and are increasingly applied in intelligent tutoring and feedback systems. The presentation addressed the research question: *"How can we characterize the AI-generated feedback if provided with a task description and a student solution as input?"*.

Several existing LLMs (e.g., GPT-3.5, GPT-4 Turbo, GPT-o1-preview, Llama3.2-3B, DeepSeek-R1-14B) were evaluated on 55 authentic student submissions from two introductory programming assignments. The feedback was qualitatively analyzed and compared (cf. [1, 2, 3, 4]).

Overall, the models are capable of providing structured, detailed, and personalized feedback, but there are several differences in terms of compliance with specifications, correctness, inconsistencies, and redundancy. Approaches to addressing some of these issues have been discussed, although some may be inherent to the LLM approach.

References

1 Azaiz, Imen; Deckarm, Oliver; Strickroth, Sven (2023, December). *AI-enhanced Auto-Correction of Programming Exercises: How Effective is GPT-3.5?*. International Journal of Engineering Pedagogy (iJEP) 8/13, pp. 67–83, DOI 10.3991/ijep.v13i8.45621.

2 Azaiz, Imen; Kiesler, Natalie; Strickroth, Sven (2024). *Feedback-Generation for Programming Exercises With GPT-4*. Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2024, Association for Computing Machinery, pp. 31–37, DOI 10.1145/3649217.3653594.

3 Azaiz, Imen; Kiesler, Natalie; Strickroth, Sven; Zhang, Anni (2025, July). *Open, Small, Rigmarole – Evaluating Llama 3.2 3B's Feedback for Programming Exercises*. International Journal of Engineering Pedagogy (iJEP) 5/15, pp. 57–73, DOI 10.3991/ijep.v15i5.55359.

4 Azaiz, Imen; Felippo, Konrad; Strickroth, Sven (2025). *Small but Competitive – Evaluating DeepSeek-R1 Among Diverse Open LLMs for Formative Programming Feedback*. In: Greubel, André; Strickroth, Sven; Striewe, Michael (eds.): Proceedings of the 7th Workshop "Automatische Bewertung von Programmieraufgaben" (ABP2025), pp. 97–106, DOI 10.18420/abp2025_10. Material available at Zenodo: .

3.2 Developer attitudes towards generative AI

Jamie Benario (Google – Chicago, US)

In this talk, I present recent research at Google and other tech companies that investigate how developers perceive generative AI tools at work. One study that we discuss is research on developer attitudes towards generative AI and their career. The research shows that developers have very complicated feelings, responding with both positive and negative attitudes towards the technology. In the rest of the presentation we discuss related research from around Google and other technology companies that provide insight into these thoughts.

3.3 Generative AI in Upper-level Computing Courses

Dennis Bouvier (United States Air Force Academy, US)

Generative AI (GenAI) is playing an increasingly influential role in computing education across all levels, offering new opportunities to support both teaching and learning. However, its effective integration raises critical concerns related to trust, academic integrity, and broader social and ethical implications. While substantial attention has been given to GenAI use in introductory programming courses (e.g., CS0/CS1), there remains a notable gap in research addressing its application in "upper-level" computing courses such as software engineering, human-computer interaction, algorithms, operating systems, and theoretical computer science.

This presentation gives a brief overview of the early work 2025 ITICSE conference Working Group 2 has drafted for a report that will present two complementary studies: a systematic literature review of GenAI interventions in upper-level computing education, and a survey of computing educators on their practices and perspectives regarding GenAI integration in these contexts.

3.4 What Does It Mean to Program in the Age of AI?

Claus Brabrand (IT University of Copenhagen, DK)

Joint work of Sebastian Nicolajsen, Claus Brabrand
Main reference Sebastian Mateos Nicolajsen, Claus Brabrand: "What Is Programming?", Commun. ACM, Vol. 68(6), pp. 28–30, 2025.
URL https://doi.org//10.1145/3713068

Introduction

Two years ago, while visiting colleagues in Uppsala, we were asked a deceptively simple question: "What does it mean to program?" Despite decades of teaching programming (CS1 and beyond), the question left us momentarily without an answer. What seemed trivial was profoundly unsettling. Today, with the rise of generative AI, the question has only grown more urgent: if programming is just "writing code," is programming itself becoming obsolete?

Historical Perspectives

Definitions of programming have expanded over time:

- **1950s:** programming as drawing up the "schedule" of operations to perform a calculation [1]. (To) program ≈ Calculation and writing a sequence of operations.
- **1970s (Knuth & Dijkstra):** programming as the *art* of composing programs, requiring aesthetics, ingenuity, and creativity [2, 3]. (To) program ≈ Creativity and composition of programs.
- **1985 (Naur):** a program's "life" depends on developers' understanding of how it supports the problem domain [4]. (To) program ∋ Understanding problem & problem domain.
- **2020s (Ko):** programming must account for societal impact: limitations, biases, and ethical considerations [5]. (To) program ∋ Consideration of unintended impact of Programs.

From calculation *to* creativity *to* understanding *to* consideration, the theoretical arc consistently well points beyond (just) "code."

Educator Perspectives

To explore how programming is understood in practice, we conducted a comprehensive study of Danish higher-education programs. Since computing is not mandatory in Danish primary or secondary school, introductory programming courses (CS1) in higher education often represent students' *first formal encounter with programming*: **the educational frontier**. By systematically identifying all such courses nationwide and surveying their educators, we obtained a representative snapshot of how programming is introduced at scale. The results revealed a sharp contrast:

- **What is programming?** educators' definitions of *programming* typically aligned with a narrow view:

  ```
  [design] -implement-> [code]
  ```

- **What is a (good) programmer?** while their definitions of a *good programmer* reflected a broad view spanning the entire problem–solution cycle (with potential iteration):

  ```
  [problem] -analyze-> [spec] -design-> [design]
    -implement-> [code] -evaluate-> [system]
  ```

While no single consensus emerged, the pattern was clear: programming itself was reduced to coding, whereas being a good programmer demanded engagement with the broader cycle.

Tension & Implications

This mismatch is consequential. Students are often implicitly taught that *programming = coding*. In the age of AI, this view invites a dangerous inference: if programming is "just coding," then programming can simply be *outsourced* to generative models. This will not work since programming is inherently more than code generation. Also, it risks depriving students of the chance to grapple with the harder, more enduring aspects of programming: analysis, specification, evaluation, and system-level reasoning.

Call to Action

Our findings echo a broader gap in AI-in-Education research. As highlighted in [6], most empirical evidence addresses the **Solution** domain. Some studies examine **Problem** → **Solution**, but still end in the Solution domain. Virtually none examine the **Problem** domain itself. None explore **Solution** → **Problem**, i.e., evaluating solutions back against the problem domain.

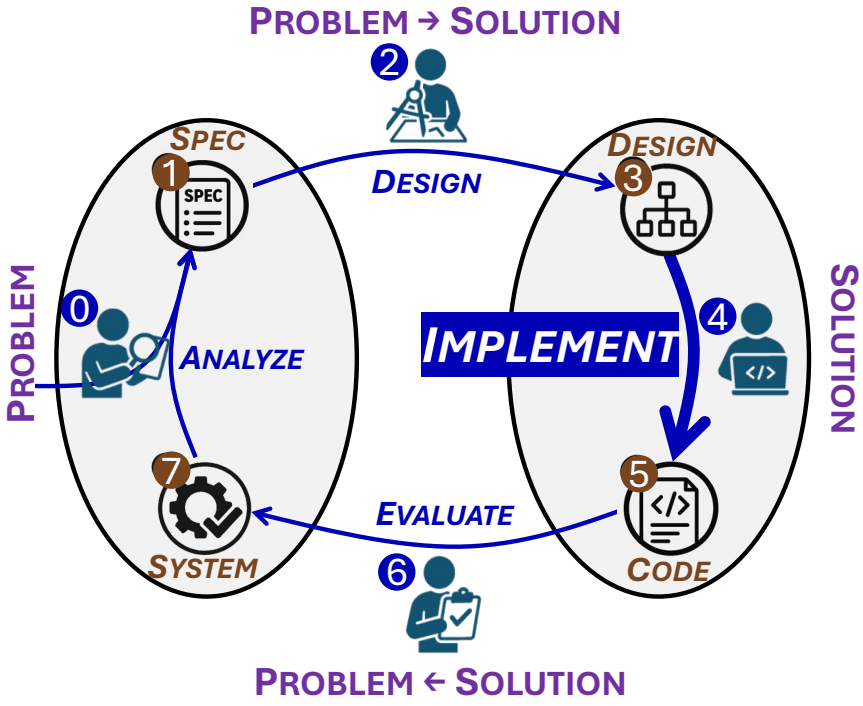


Figure 1 Educators' definitions: narrow vs. broad views of programming.

The way forward is clear. Both education and research must shift attention toward the **Problem domain**:

- **Analyze:** cultivating deep understanding of the problem.
- **Evaluate:** ensuring solutions actually address it.

These stages are hardest to outsource to AI and most crucial for developing durable programming competence.

Conclusion

We do not argue for a single hegemonic definition of programming. Instead, we call for an explicit conversation – with colleagues, with students, and with society. Programming cannot be reduced to the narrow slice of coding. It encompasses the broader, intangible qualities of problem analysis, domain understanding, evaluation, and societal consideration – the very human qualities that will remain vital, probably even more so, even in the age of AI.

References

1 A. F. Blackwell. What is programming? In *Proceedings of the 14th Workshop of the Psychology of Programming*, page 14. Citeseer, 2002.
2 D. E. Knuth. Computer programming as an art. In *ACM Turing Award Lectures*, 1974.
3 E. W. Dijkstra. A short introduction to the art of programming. Vol. 4. Technische Hogeschool, Eindhoven, 1971.
4 P. Naur. Programming as theory building. *Microprocessing and Microprogramming*, 15(5), 1985.
5 A. J. Ko et al. It is time for more critical CS education. *Communications of the ACM*, 63(11), Nov. 2020.

6 J. Prather, J. Leinonen, N. Kiesler, J. G. Benario, S. Lau, S. MacNeil, N. Norouzi, S. Opel, V. Pettit, L. Porter, B. N. Reeves, J. Savelka, D. H. Smith IV, S. Strickroth, and D. Zingaro. Beyond the hype: A comprehensive review of current trends in generative AI research, teaching practices, and tools. In *Proc. ITiCSE-WGR 2024*, pp. 1–39, Milan, Italy, 2024. doi:10.1145/3689187.3709614.

3.5 Howzat? Appealing to Expert Judgement for Evaluating Human and AI Next-Step Hints for Novice Programmers

Neil Brown (King's College London, GB)

Joint work of Neil Brown, Paul Denny, Juho Leinonen
Main reference Neil C. C. Brown, Pierre Weill-Tessier, Juho Leinonen, Paul Denny, Michael Kölling: "Howzat? Appealing to Expert Judgement for Evaluating Human and AI Next-Step Hints for Novice Programmers", ACM Trans. Comput. Educ., Vol. 25(3), pp. 1–43, 2025.
URL https://doi.org//10.1145/3737885

Motivation: Students learning to program often reach states where they are stuck and can make no forward progress – but this may be outside the classroom where no instructor is available to help. In this situation, an automatically generated next-step hint can help them make forward progress and support their learning. It is important to know what makes a good hint or a bad hint, and how to generate good hints automatically in novice programming tools, for example using Large Language Models (LLMs).

Method and participants: We recruited 44 Java educators from around the world to participate in an online study. We used a set of real student code states as hint-generation scenarios. Participants used a technique known as comparative judgement to rank a set of candidate next-step Java hints, which were generated by Large Language Models (LLMs) and by five human experienced educators. Participants ranked the hints without being told how they were generated. The hints were generated with no explicit detail given to the LLMs/humans on what the target task was. Participants then filled in a survey with follow-up questions. The ranks of the hints were analysed against a set of extracted hint characteristics using a random forest approach.

Findings: We found that LLMs had considerable variation in generating high quality next-step hints for programming novices, with GPT-4 outperforming other models tested. When used with a well-designed prompt, GPT-4 outperformed human experts in generating pedagogically valuable hints. A multi-stage prompt was the most effective LLM prompt. According to a fitted random forest model, the two most important factors of a good hint were length (80–160 words being best), and reading level (US grade nine or below being best). Offering alternative approaches to solving the problem was considered bad, and we found no effect of sentiment.

Conclusions: Automatic generation of these hints is immediately viable, given that LLMs outperformed humans – even when the students' task is unknown. Hint length and reading level were more important than several pedagogical features of hints. The fact that it took a group of experts several rounds of experimentation and refinement to design a prompt that achieves this outcome suggests that students on their own are unlikely to be able to produce the same benefit. The prompting task, therefore, should be embedded in an expert-designed tool.

3.6 Opening Remarks: Generative AI and the Future of Programming Education

Paul Denny (University of Auckland, NZ)

Joint work of Natalie Kiesler, Michelle Craig, James Prather, Paul Denny

The seminar opened with reflections on the rapid and transformative developments in Generative AI and their implications for programming education. This is an area marked by both excitement and uncertainty, with leaders from academia and industry expressing diverging views. These range from predictions that traditional university education, particularly in technical disciplines, may become obsolete, to optimism that AI will enhance learning and professional productivity. These perspectives were noted as mirroring recent debates in Communications of the ACM, where opposing viewpoints have argued either for the end of programming as it is currently understood or for a new era of AI-augmented programming practice.

The organisers of the seminar, Natalie Kiesler, Michelle Craig, James Prather, and Paul Denny, were excited at the prospect of fostering collaboration among researchers and practitioners from computing education, software engineering, HCI, and related fields. Participants were invited to explore how Generative AI is reshaping the ways that programming is taught, and how pedagogical approaches, assessment practices, and curricula need to evolve to ensure that AI supports rather than replaces meaningful learning in programming.

The organisers paid a special tribute to Dr Brett Becker, whose early work and vision were instrumental in the establishment of this seminar. Dr Becker, who passed away in October 2024, was remembered with deep appreciation for his contributions to the field and his role in bringing this community together.

3.7 Never in my Wildest Dreams: GenAI Agent-Based Software Development

Christopher D. Hundhausen (Oregon State University – Corvallis, US)

URL http://tiny.cc/GenAIAgents

Agentic GenAI tools, such as Copilot Agent, have been recently integrated into IDEs such as Visual Studio Code. Using thse tools, software developers can orchestrate software development by describing software requirements to a GenAI agent in plain English. The agent then implements the requirements by directly modifying a codebase. This process resembles a computational steering model with a human in the loop to monitor development progress and steer the agent toward the desired solution. I reflect on one month of intensive use of a GenAI agent, identifying high-level abstractions that a developer uses in lieu of computer code. I propose a set of general student learning outcomes that computing educators might use to design and learning activities to facilitate this form of software development and assess student learning. A video demo of my interactions and reflections can be found at `http://tiny.cc/GenAIAgents`.

3.8 Edugator: An AI-enabled Tool for Creating and Delivering Interactive Computing Content

Amanpreet Kapoor (University of Florida – Gainesville, US)

Joint work of Marc Diaz, Dustin Karp, Prayuj Tuli, Amanpreet Kapoor
Main reference Marc Diaz, Dustin Karp, Prayuj Tuli, Amanpreet Kapoor: "Edugator: An AI-enabled Tool for Creating and Delivering Interactive Computing Content", in Proc. of the 56th ACM Technical Symposium on Computer Science Education V. 2, SIGCSE TS 2025, Pittsburgh, PA, USA, 26 February 2025 – 1 March 2025, p. 1732, ACM, 2025.
URL https://doi.org//10.1145/3641555.3705025

Edugator is a browser-based, AI-enabled tool designed to help instructors of introductory computing courses create and deliver interactive educational content. It streamlines the content authoring process by incorporating generative AI models into both the creation and delivery stages. Using this tool, instructors can create bespoke interactive computing lessons and programming problems by providing a prompt and a few clicks. They can also author templates and test cases in programming languages such as C++, Java, C, and Python. Additionally, instructors can validate programming problems by running them against an auto-generated solution, allowing them to refine the problems before releasing it to students, preventing misinformation or ambiguity. Students can complete lessons and solve programming problems in a browser-based text editor receiving immediate feedback. They can also interact with a large language model-powered AI chatbot that scaffolds a student on how to approach the problem without giving out solutions. Edugator is built using modern web frameworks and the goal of the tool is to accelerate the adoption of automated assessment tools by minimizing the challenges instructors face with such tools. It also supports Learning Tools Interoperability (LTI), allowing seamless integration with learning management systems (LMS). Tools like Edugator streamline assessment authoring and delivery for Instructors while supporting student learning by promoting learning-by-doing and providing meaningful, personalized feedback. More information about the tool can be found at `https://edugator.app/`.

3.9 Inevitable AI? Reconsidering the "Inevitable" Integration of Generative AI in Computing Education

Hieke Keuning (Utrecht University, NL)

One of the objectives of this Dagstuhl Seminar, and of many other initiatives on Generative AI, is to develop recommendations for the pedagogical integration of GenAI in programming courses. I aim to challenge this seemingly "inevitable" integration.[1]

Large Language Models are built on human labor exploitation, stolen data, and are detrimental to the environment [2]. Unfortunately, this is usually an afterthought in many papers and talks. It is impossible to ethically justify their use at such a large scale. A huge amount of money is involved, and companies' interests are clearly not in human learning

[1] https://leonfurze.com/2025/04/28/the-myth-of-inevitable-ai/

[11]. The hype is unprecedented. The main goal of the major AI companies is automation, ultimately replacing much of human work. Turning to the consequences of its use, there is evidence of reduced critical thinking when knowledge workers use GenAI [12]. The benefits for programmers are debatable[2]; a recent study has shown that AI makes developers slower, although they thought it was the opposite [9].

So is the promise of revolutionizing education false? Current meta-reviews are not convincing [3], and many claims are unsupported by evidence [4]. Teachers spend a lot of time "reviewing, repairing and sometimes completely reworking AI-produced outputs" [8]. In computing education, researchers have observed a "widening gap" between weaker and stronger students [1] and signs of "social erosion" because students no longer ask each other for help [5]. While many new tools have emerged, they are often not built on known pedagogy or learning theory [7].

In the end, do we decide, or the technology [13]? "AI is unavoidable, but not inevitable".[3] Pretending it is not there does not make any sense, but we can take the lead in deciding what to do with it. We can take a step back and resist the hype [6]. Put pedagogy first, provide guardrails and scaffolding for authentic learning experiences. When building tools, only use LLMs for specific use cases, combining them with proven techniques (e.g. [10]). Finally, let us go back to the real problems at hand, and focus on community and learning.

References

1. Prather, J., Reeves, B., Leinonen, J., MacNeil, S., Randrianasolo, A., Becker, B., Kimmel, B., Wright, J. & Briggs, B. The Widening Gap: The Benefits and Harms of Generative AI for Novice Programmers. *Proceedings Of The 2024 ACM Conference On International Computing Education Research*. pp. 469-486 (2024)
2. Muldoon, J., Graham, M. & Cant, C. Feeding the machine: the hidden human labour powering AI. (Canongate Books,2024)
3. Weidlich, J., Gašević, D., Drachsler, H. & Kirschner, P. ChatGPT in Education: An Effect in Search of a Cause. *Journal Of Computer Assisted Learning*. **41** (2025)
4. Kohn, T. From Imitation Games to Robot-Teachers: A Review and Discussion of the Role of LLMs in Computing Education. *Journal Of Computer Assisted Learning*. **41** (2025)
5. Hou, I., Man, O., Hamilton, K., Muthusekaran, S., Johnykutty, J., Zadeh, L. & MacNeil, S. 'All Roads Lead to ChatGPT': How Generative AI is Eroding Social Interactions and Student Learning Communities. *Proceedings Of The 30th ACM Conference On Innovation And Technology In Computer Science Education*. pp. 79-85 (2025)
6. Rudolph, J., Ismail, F., Tan, S. & Seah, P. Don't believe the hype. AI myths and the need for a critical approach in higher education. *Journal Of Applied Learning And Teaching*. **8**, 6-27 (2025)
7. Topali, P., Haelermans, C., Molenaar, I. & Segers, E. Pedagogical considerations in the automation era: A systematic literature review of AIEd in K-12 authentic settings. *British Educational Research Journal*. (2025)
8. Selwyn, N., Ljungqvist, M. & Sonesson, A. When the prompting stops: exploring teachers' work around the educational frailties of generative AI tools. *Learning, Media And Technology*. pp. 1-14 (2025)
9. Becker, J., Rush, N., Barnes, E. & Rein, D. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. *ArXiv Preprint ArXiv:2507.09089*. (2025)

[2] https://blog.glyph.im/2025/06/i-think-im-done-thinking-about-genai-for-now.html
[3] https://marcwatkins.substack.com/p/ai-is-unavoidable-not-inevitable

10 Birillo, A., Artser, E., Potriasaeva, A., Vlasov, I., Dzialets, K., Golubev, Y., Gerasimov, I., Keuning, H. & Bryksin, T. One step at a time: Combining llms and static analysis to generate next-step hints for programming tasks. *Proceedings Of The 24th Koli Calling International Conference On Computing Education Research*. pp. 1-12 (2024)

11 Olson, P. Supremacy: AI, ChatGPT, and the Race that Will Change the World. (St. Martin's Press,2024)

12 Lee, H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R. & Wilson, N. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings Of The 2025 CHI Conference On Human Factors In Computing Systems*. pp. 1-22 (2025)

13 Padiyath, A. Do I Have a Say in This, or Has ChatGPT Already Decided for Me?. *XRDS*. pp. 52-55 (2024)

3.10 Beyond the Hype: A Sneak Peek into the Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools

Natalie Kiesler (Technische Hochschule Nürnberg, DE) and James Prather (Abilene Christian University, US)

Joint work of James Prather, Juho Leinonen, Natalie Kiesler, Jamie Gorson Benario, Sam Lau, Stephen MacNeil, Narges Norouzi, Simone Opel, Vee Pettit, Leo Porter, Brent N. Reeves, Jaromír Savelka, David H. Smith, Sven Strickroth, Daniel Zingaro

Main reference James Prather, Juho Leinonen, Natalie Kiesler, Jamie Gorson Benario, Sam Lau, Stephen MacNeil, Narges Norouzi, Simone Opel, Vee Pettit, Leo Porter, Brent N. Reeves, Jaromír Savelka, David H. Smith, Sven Strickroth, Daniel Zingaro: "Beyond the Hype: A Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools", in Proc. of the 2024 Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE 2024, Milan, Italy, 8 July 2024, pp. 300–338, ACM, 2024.

URL https://doi.org//10.1145/3689187.3709614

Generative AI (GenAI) keeps advancing rapidly, and the literature in computing education is expanding almost as quickly. Initial responses to GenAI tools were mixed between panic and utopian optimism. Many were quick to point out the opportunities and challenges of GenAI. Researchers reported that these new tools are capable of solving most introductory programming tasks and are causing disruptions throughout the curriculum. These tools can write and explain code, enhance error messages, create resources for instructors, and even provide feedback. In 2024, new research started to emerge, focusing on the effects of GenAI usage in the computing classroom. At the same time, a new class of tools is being developed that can provide personalized feedback to students on their programming assignments or teach both programming and prompting skills. With the literature expanding so rapidly, an ITiCSE working group aimed to summarize and explain what is happening on the ground in computing classrooms. In our lightning talk presentation, we provided the results of a systematic literature review; a survey of educators and industry professionals; and interviews with educators using GenAI in their courses, educators studying GenAI, and researchers who create GenAI tools to support computing education. The triangulation of these methods and data sources helps expand the understanding of GenAI usage and perceptions at this critical moment for our community.

3.11 Evidence of Learning Loss and Teaching Fundamentals

Colleen Lewis (University of Illinois Urbana-Champaign, US)

Joint work of Binglin Chen, Colleen M. Lewis, Matthew West, Craig Zilles
Main reference Binglin Chen, Colleen M. Lewis, Matthew West, Craig Zilles: "Plagiarism in the Age of Generative AI: Cheating Method Change and Learning Loss in an Intro to CS Course", in Proc. of the Eleventh ACM Conference on Learning @ Scale, L@S '24, p. 75–85, Association for Computing Machinery, 2024.
URL https://doi.org//10.1145/3657604.3662046

In my lightning talk I discussed two projects that are relevant to the teaching and learning of computer science (CS) in the age of Generative AI. The first project looks at learning loss from student plagiarism within an introductory programming course [1]. We collected data in an intro Python course for non-CS majors before and after the wide availability of ChatGPT. We developed four indicators of plagiarism. We found that these indicators appearing on homework submissions were correlated with lower performance on a final exam taken within a testing facility, while controlling for a student's performance on an exam early in the semester. This speaks to the ways in which plagiarism appears to decrease learning and the age of Generative AI provides easy opportunities for plagiarism. The second project looks at using physical objects to help students reason about foundational ideas within programming [2, 3]. This includes variable types, variable assignment, and function calls. This work builds on taken for granted practices within mathematics education of using physical objects to help students learn about quantity. Resources for this project are available at CSTeachingTIps.org/3D.

References

1 Chen, B., Lewis, C. M., West, M., & Zilles, C. (2024). Plagiarism in the Age of Generative AI: Cheating Method Change and Learning Loss in an Intro to CS Course. Learning at Scale 2024. https://doi.org/10.1145/3657604.36620
2 Lewis, C. M. (2021). Physical Java Memory Models: A Notional Machine. ACM SIGCSE Proceedings. 52(1). https://doi.org/10.1145/3408877.3432477
3 Lewis, C. M., Hernandez, M., Kuo, A., McDowell, H., Roller, H. (2025). Experience Report: Physical Models of Java Inheritance. SIGCSE 2025. Pittsburg, PA. 10.1145/3641554.3701871

3.12 Disciplinary Identity and Design Methods for GenAI

Kevin Lin (University of Washington – Seattle, US)

Joint work of Kevin Lin, Alannah Oleson, Anna Batra, Iris Zhou, Suh Young Choi, Chongjiu Gao, Yanbing Xiao, Sonia Fereidooni

One promise of generative AI for programming is that it can help us build software by quickly translating specifications into implementations. If we follow the premise, then skills like analyzing the qualities of specifications and evaluating the correctness of implementations could be more important to emphasize. But this presumes we know what software we want to build in the first place. Our choices shape not only what skills students learn but also their disciplinary values interpretation: "a process by which students reflect on the values of a disciplinary domain, as well as who they are and might become in relation to the domain" [1].

My scholarship has explored redesigning technologies as a way to shape disciplinary values interpretation by drawing on design methods such as iterative design [2, 3]. Iterative design is a software development practice that involves prototyping, testing, analyzing, and refining technology. Instead of assigning students a complete specification of a program to implement and focusing on evaluating the qualities of the final product, students could showcase their software development process over time. We can then ask questions about each step of the process.

Generative AI is often framed as a productivity tool, but what does productivity free us to do? Teaching students design methods could empower them to ask bigger questions about their work and challenge them to reflect on what exactly they hope to achieve in their future computing careers [4].

References

1 Sepehr Vakil. 2020. "I've Always Been Scared That Someday I'm Going to Sell Out": Exploring the relationship between Political Identity and Learning in Computer Science Education. In Cognition and Instruction, 38(2), 87–115. https://doi.org/10.1080/07370008.2020.1730374
2 Alannah Oleson, Meron Solomon, Christopher Perdriau, Amy Ko. 2023. Teaching Inclusive Design Skills with the CIDER Assumption Elicitation Technique. ACM Trans. Comput.-Hum. Interact. 30, 1, Article 6 (February 2023), 49 pages. https://doi.org/10.1145/3549074
3 Kevin Lin. 2024. An Invitation to Reimagine: Empowering Students to Redesign Computing Problems and Artifacts. Invited talk. https://kevinl.info/an-invitation-to-reimagine/
4 Anna Batra, Iris Zhou, Suh Young Choi, Chongjiu Gao, Yanbing Xiao, Sonia Fereidooni, Kevin Lin. 2024. "It Can Relate to Real Lives": Attitudes and Expectations in Justice-Centered Data Structures & Algorithms for Non-Majors. In Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024). Association for Computing Machinery, New York, NY, USA, 88–94. https://doi.org/10.1145/3626252.3630754

3.13 APFEL – Adaptive Programming Feedback for E-Learning

Dominic Lohr (Universität Erlangen-Nürnberg, DE)

Main reference Dominic Lohr, Marc Berges, Abhishek Chugh, Michael Striewe: "Adaptive Learning Systems in Programming Education: A Prototype for Enhanced Formative Feedback", in Proc. of the DELFI 2024 - Die 22. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V., DELFI 2024, Fulda, Germany, September 9-11, 2024, LNI, Vol. P-356, Gesellschaft für Informatik e.V., 2024.
URL https://doi.org//10.18420/DELFI2024_57

In my talk, I presented the tool APFEL, an adaptive learning environment for programming education that integrated GenAI in a controlled and transparent wa.

Unlike systems that leave it to LLMs to "decide" what kind of feedback to generate or provide and *when*, APFEL ensures that the feedback process is traceable and understandable for learners – if they want to know "Why did I get this feedback here?"

At its core, APFEL uses a symbolic/rule-based layer to determine the type of feedback that should be provided, based on empirical findings and pedagogical concepts and principles. This context is then used to systematically assemble a prompt, which is further enriched with additional context from a domain model and a learner model using RAG. This ensures that the resulting feedback is personalized and also didactically grounded and traceable.

3.14 Who is the author?

Andrew James Luxton-Reilly (University of Auckland, NZ)

Humans are considered to be the agents responsible for actions when they use tools to complete those actions. When we say "I cut down a tree", we attribute the act to the human rather than the axe. Similarly, when we create code using a tool such as GenAI, we should attribute the act of creation to the human rather than the machine.

Although this view challenges some academics, publishers such as ACM have already decided that only humans can be authors, and they are responsible for the product of GenAI use.

We conclude that users of GenAI tools are the authors of the output produced. Since these tools are ubiquitous, authentic assessment tasks should allow students to use GenAI, and the work product should be graded in the same manner as any other student work. To operationalise this approach, we adopt a two-lane approach in which we make a distinction between secure and insecure assessments.

For secure assessment, we may prohibit the use of GenAI to determine what the student can achieve independent of tools. For insecure assessment, students are permitted to use GenAI and must be graded with the view that they are the author, and are responsible for the product that they produce using tools.

3.15 All Roads Lead to ChatGPT: The Negative Impacts on Learning Communities

Stephen MacNeil (Temple University – Philadelphia, US)

Joint work of Irene Hou, Owen Man, Kate Hamilton, Srishty Muthusekaran, Jeffin Johnykutty, Leili Zadeh, Stephen MacNeil
Main reference Irene Hou, Owen Man, Kate Hamilton, Srishty Muthusekaran, Jeffin Johnykutty, Leili Zadeh, Stephen MacNeil: "'All Roads Lead to ChatGPT': How Generative AI is Eroding Social Interactions and Student Learning Communities", in Proc. of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2025, Nijmegen, The Netherlands, 27 June 2025 - 2 July 2025, pp. 79–85, ACM, 2025.
URL https://doi.org//10.1145/3724363.3729024

Generative AI provides students with valuable support even late at night when their friends are asleep. However, as students rely more heavily on these tools for help, instead of their friends, what will be the impact? Based on semi-structured interviews with 17 undergraduate computing students, we found that help-seeking requests are now often mediated by generative AI, with students frequently redirecting questions from their peers to generative AI instead of providing assistance themselves. This has the potential to undermine relationships and create a transactional classroom environment where friends never have a chance to form and learning communities erode. Students also reported feeling increasingly isolated from their peers and demotivated as a result. Our work is call to action to refocus our efforts on fostering vibrant learning communities and centering social interactions in an age of automation.

3.16 Prompt Programming

Victor-Alexandru Padurean (MPI für Software Systems – Saarbrücken, DE)

Joint work of Victor-Alexandru Padurean, Paul Denny, Alkis Gotovos, Adish Singla
Main reference Victor-Alexandru Padurean, Paul Denny, Alkis Gotovos, Adish Singla: "Prompt Programming: A Platform for Dialogue-based Computational Problem Solving with Generative AI Models", in Proc. of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2025, Nijmegen, The Netherlands, 27 June 2025 - 2 July 2025, pp. 458–464, ACM, 2025.
URL https://doi.org//10.1145/3724363.3729094

Computing students increasingly rely on generative AI tools for programming assistance, often without formal instruction or guidance. This highlights a need to teach students how to effectively interact with AI models, particularly through natural language prompts, to generate and critically evaluate code for solving computational tasks.

To address this, we developed a novel platform for prompt programming that enables authentic dialogue-based interactions, supports problems involving multiple interdependent functions, allows manual code edits of generated code, and offers on-request execution of generated code.

Data analysis from students using the tool in introductory programming courses shows high engagement, with most prompts occurring in multi-turn dialogues. Furthermore, students were highly selective about the code they chose to execute. Reflection data also indicates that the majority of students preferred a balanced mix of natural language prompting and manual code editing when tackling harder problems.

3.17 Learn AI-Assisted Python Programming

Leo Porter (University of California – San Diego, US) and Daniel Zingaro (University of Toronto Mississauga, CA)

The advent of Generative AI tools is rapidly transforming the landscape of software development, presenting both challenges and opportunities for computer science education. In this talk, we address the critical need to adapt introductory programming curricula to this new reality. We share the journey of creating the textbook "Learn AI-Assisted Python Programming" and developing a corresponding CS1 course that embraces AI tools. The core of our approach is to shift the focus from traditional coding syntax to a new set of essential skills for the modern developer, including prompt engineering, critical code evaluation, testing, debugging, and problem decomposition in an AI-assisted environment.

We provide a detailed account of our pilot "CS1-LLM" course at UC San Diego, which served over 500 students. We discuss the course's design, which featured open-ended, creative projects across various domains to keep students engaged. We conclude with practical lessons learned from our teaching experience and an overview of our ongoing work to support educators and build a broader community around integrating AI into computer science education.

3.18 A New Curriculum for Computer Science that integrated GenAI: A Collaboration between Google and Academia

James Prather (Abilene Christian University, US)

Google's Tech Exchange program has been shaping and building engineers for many years. The advent of GenAI caused some panic among the creators, and it became a critical question about what to do. It was decided that the curriculum must be entirely reimagined. Google partnered with over a dozen universities to rework its curriculum to be GenAI forward from the very beginning. The redesigned courses were CS1, CS2, Applied Algorithms, Software Engineering, and Project Management. The results show that students engaged, worked hard, and learned traditional coding and computer science concepts. Students also increased their general self-efficacy over the test semester as well as their GenAI-specific self-efficacy. We believe this curriculum represents an effective first step in redesigning our courses to account for this fast-growing technology.

3.19 Experiences from an AI Task Force at a Large Institution

Karen Reid (University of Toronto, CA)

Joint work of The University of Toronto Task Force on Generative AI
Main reference The University of Toronto Task Force on Generative AI: "Report: Toward an AI-ready university – University of Toronto." (June 2025). Retrieved September 19, 2025.
URL https://ai.utoronto.ca/u-of-t-ai-task-force/report/

The University of Toronto Task Force on AI explored the impact of generative AI on all aspects of the University. I served on the main task force and co-chaired the Teaching and Learning Working group. During our consultations we heard many faculty despairing that unsupervised assessments no longer promote student development and learning since students can use GenAI tools to complete much of the work. The working group report highlights the need for GenAI literacy programming for both faculty and students, the importance of reconsidering learning outcomes across the curriculum, the challenge of providing formative feedback at scale, and noted that classroom dynamics are shifting as students and faculty alike bring their AI avatars to class with them. The task force reports recommend that the university invest in AI technology so that faculty and staff may experiment with and develop tools to support student learning. As we refined recommendations, we kept coming back to the idea that "good teaching is still good teaching." In other words, best teaching practices remain resilient in the age of GenAI. Ultimately, the task force focused on the university as an "institution dedicated to the development of human potential." Key takeaways include the need for greater transparency with students about how learning happens and a renewed determination to foster social interaction among students to enhance their learning.

3.20 AISOP – Using AI to Leverage E-Portfolios in Teaching

Daniel Schiffner (DIPF – Frankfurt am Main, DE)

Joint work of Daniel Schiffner, Wolfgagng Müller
Main reference Alexander Gantikow, Andreas Isking, Paul Libbrecht, Wolfgang Müller, Sandra Rebholz: "On the Creation of Classifiers to Support Assessment of E-Portfolios", in Proc. of the IEEE International Symposium on Multimedia, ISM 2023, Laguna Hills, CA, USA, December 11-13, 2023, pp. 297–302, IEEE, 2023.
URL https://doi.org//10.1109/ISM59092.2023.00057

In order to make the learning process visible, E-Portfolios have been utilized for many years. However, for both, teachers and students, this is considered an time-consuming and tedious task. As part of the project AISOP at the PH-Weingarten, these tasks have been simplified using AI methodologies.

Teachers define a concept map, which the tool uses to evaluate the created portfolios. Students get feedback on their current state before going into the examination, which only bases on the portfolio.

As code, it looks for the student like this:

```
while(learning) {
  write_portfolio() //here reflection with learning happens
  gather_experience()
  update_and_evaluate_using_ai() // concept map and AI tools are triggered
  get_feedback()
}
prepare_for_exam()
do_exam()
```

For the teachers, it could be represented as follows:

```
define_concept_map()
while(learning) {
  assign_tasks()
  gather_insights_from_ai() //using dashboards, etc.
  provide_additional_feedback() // human in the loop
}
check_portfolios_with_ai()
//prior to examination, check for missing entries, unclear documentation, etc.
do_exam_using_portfolio()
```

The project was defined and started in the pre-GPT era, with ChatGPT disrupting a lot of development. Many possible improvements exist, as higher quality evaluation is now possible. Currently, the project is being finished, and the idea pushed and feedback on it's implementation as a general tool collected. We hope that the tool gets the opportunity to be finished and provided to other teachers. This will facilitate a different type of assessment that cannot be simply solved by memorization and and requires some time and effort for students. The time spent helps students to better understand larger concepts and emphasizes continuous learning

3.21 Generative AI in Software Engineering

Titus Winters (Adobe – New York, US)

A summary of recent trends, stats, and beliefs regarding the use of generative AI within the tech industry.

3.22 Accessibility of GenAI Tools with Screen Readers

Daniel Zingaro (University of Toronto Mississauga, CA)

In this talk, I discuss two separate GenAI accessibility topics.

First, that there's an unequivocal pro of GenAI tools (in all of the needed pro-con discussions!): GenAI-powered tools are making visual-related accessibility improvements that I never thought possible. They are describing videos, making inaccessible apps accessible, helping with usability concerns on websites, describing graphs and other diagrams from papers, etc. Many of these challenges have been understood but not solved for decades ("add alt tags to your images!"); GenAI is allowing us to go beyond a mere description of visual material to full-on interaction.

Second, that we as a community need to ensure that the GenAI tools we ask our students are indeed accessible! In particular, I described the inherent accessibility of commandline-based GenAI tools that otherwise maintain a user's existing workflow (rather than requiring them to use IDEs that may take weeks to learn and that themselves may not be accessible).

3.23 Human-AI Interaction Challenge: How to Ensure Continued Growth of a Human as an Expert?

Jaromír Šavelka (Carnegie Mellon University – Pittsburgh, US)

Traditional software development (SD) involved minimal AI and a lot of code writing and reading. While AI supported SD has been characterized by the use of manually triggered AI to do some of the code writing the recently emerged agentic SD enables AI to handle complex tasks such as planning, testing, and refactoring. Human developers still play a crucial role in defining goals, reviewing, fixing outcomes, and guiding the process. When the agent-generated codebase falls short on, e.g., functional or stylistic requirements. One can ask the agents to modify something. One can even suggest how to go about it (e.g., where in the codebase; what library to use). And one can iterate ... At some point, one may give up and modify the code themselves. Hence, (1) there will likely be less natural opportunities for one to directly write and read code; (2) a developer will likely be responsible for handling increasingly larger and complex codebases; (3) the same or even greater expertise in software engineering, including reading and writing code, might be needed. Then, how can we ensure

that future software engineers go about their everyday work in a way that systematically hones their expertise? In conclusion, there appears to be a fundamental human-AI interaction challenge: How do we design interactions between agentic SD frameworks and humans to ensure continued growth of a human as an expert?

4 Working groups

4.1 "Andy's Axe" as a Guiding Principle for our Stance on AI-in-Education (for Computing)

Ibrahim Albluwi (Princess Sumaya University for Technology – Amman, JO), Dennis Bouvier (United States Air Force Academy, US), Claus Brabrand (IT University of Copenhagen, DK), Michelle Craig (University of Toronto, CA), Rodrigo Duran (Federal Institute of Brasília, BR), Christopher D. Hundhausen (Oregon State University – Corvallis, US), Kevin Lin (University of Washington – Seattle, US), Andrew James Luxton-Reilly (University of Auckland, NZ), Leo Porter (University of California – San Diego, US), Karen Reid (University of Toronto, CA), David H. Smith IV (Virginia Polytechnic Institute – Blacksburg, US), Claudia Szabo (University of Adelaide, AU), Shubbhi Taneja (Worcester Polytechnic Institute, US), Michel Wermelinger (The Open University – Milton Keynes, GB), Titus Winters (Adobe – New York, US), and Daniel Zingaro (University of Toronto Mississauga, CA)

One of the many outcomes of Dagstuhl Seminar 25311 is the following collective *educational stance* on how to appropriately navigate the "AI-in-Education Crisis" (as it has been dubbed).

Historically, technological innovation has always preceded ethics and legislation. We argue the metaphor of *"Andy's Axe"* serves as a guiding principle informing the future educational stance on GenAI. As the analogy goes: We say: "*Andy* cut down the tree." We don't say: "An *Axe* cut down the tree." Nor do we say that the tree was cut as a result of an *"Andy–Axe collaboration."* Pressed for more detail, we would say that Andy cut down the tree, *using* the Axe. Andy's *Axe* is just another *tool*; although one that commands our attention. *Andy* is the one responsible for: (1) [in]appropriate *use* of the tool; *and* (2) [un]intended *consequences* of their actions (whether using the tool or not).

Generative Artificial Intelligence (GenAI) is here, students will use it – whether we (attempt to) ban it or not – and they will need it in their future careers as professionals in the software industry. The alternative is educational institutions "forcing" students through *irrelevant programmes* and, ultimately, "producing" *irrelevant graduates* for the AI-empowered industry. Hence, we should not deprive future generations of students access to, and competences in, *effective*, *critical*, and *ethical use* of AI. Assuming this educational stance – *Andy* is responsible for Andy's *use* of Andy's *Axe* – the following picture emerges on the (event) horizon:

Let us *use* AI for *tedious lower-level* code writing under human guidance (the human wielding the metaphorical axe). This frees the human (student or professional) to focus more on *higher-level issues such as: problem solving*, *design*, *ethics*, *fairness*, *privacy*, *safety*, *security*, *usability*, *modularity*, *effectiveness*, *sustainability*, *trustworthiness*, and overall software *quality*.

If the adoption of AI, in any way, mirrors the adoption of the personal computer, what may come is a *rise*, not in *productivity*, but in *quality*. In any case, for this to work, we need to teach our future students how to *effectively*, *critically*, and *ethically* wield the Axe: GenAI.

4.2 GenAI in Programming Education: Hypes, Hoaxes, and Hopes

Carolin Hahnel (Ruhr-Universität Bochum, DE), Jamie Gorson Benario, Hieke Keuning (Utrecht University, NL), Natalie Kiesler (Technische Hochschule Nürnberg, DE), Tobias Kohn (KIT – Karlsruher Institut für Technologie, DE), Dennis Komm (ETH Zürich, CH), Colleen Lewis (University of Illinois Urbana-Champaign, US), Dominic Lohr (Universität Erlangen-Nürnberg, DE), Brent Reeves (Abilene Christian University, US), Jaromír Šavelka (Carnegie Mellon University – Pittsburgh, US), Jacqueline Staub (Universität Trier, DE), and Christina Weers (Goethe-Universität – Frankfurt am Main, DE)

Discussed Problems

The general theme of the working group centered on the question of what drives the hype surrounding the use of generative AI (GenAI) for educational and occupational purposes, and whether GenAI will actually deliver on its promises [3]. Two of these promises concern the future efficiency of writing code and GenAI's positive impact on learning processes and outcomes, both of which are possibly overly optimistic. We identified several strands requiring further research: (1) the effect of GenAI use on critical learning opportunities; (2) the pedagogy behind tool use and the potential gains and losses associated with its implementation; (3) the impact of GenAI use on personal and professional development; and (4) the ethical implications of GenAI use for educational purposes.

Strand 1. How does the use of GenAI change critical learning opportunities?

The concept of *desirable difficulties* in learning processes refers to challenges that can initially be effortful and frustrating but ultimately enhance understanding and personal growth [1]. The use of GenAI for learning raises questions about whether it reduces these valuable struggles. For instance, over-reliance on GenAI could create an illusion of competence, causing learners to believe they understand a topic simply because they can generate a response without truly comprehending the underlying concepts. Consequently, GenAI use may benefit advanced learners and experts by providing quick access to information, while novices may require the struggle to develop foundational understanding and critical thinking skills. Thus, these higher-order thinking skills will play a more central role in an era in which GenAI is ubiquitous.

Education may need to shift further toward teaching students how and why to learn and how to apply their knowledge effectively. This may require rethinking learning goals. In general, the focus may need to shift "from the destination to the journey." For example, self-driving cars may eventually make learning to drive obsolete. But if the same were true for, say, "reasoning," should it consequently be excluded from what we consider worth learning altogether? It is essential to define what constitutes a valuable outcome, whether personal growth, societal benefit, or environmental impact.

More research is needed to determine the concrete advantages of using GenAI in education. For example, anecdotal and empirical evidence suggest that students appreciate GenAI for providing starting points or templates [4], which can help overcome the initial hurdle of beginning to work on a task. In contrast, concern may remain that GenAI use could reduce potential creative benefits of being bored or "facing a blank page," both of which may stimulate innovative and creative thinking [5]. Overall, we need to consider how AI practices and newly developed GenAI tools relate to educational theory (see also a recent article by Hazzan and Erez [2]).

Strand 2. What do we gain and lose by using GenAI in learning processes?

Using GenAI tools may enhance but also diminish certain skills. It is crucial to quantify which skills are developed or lost through GenAI use in order to achieve a deeper understanding of how these tools can be leveraged or should be avoided. Choosing a *human-value-first* approach towards GenAI use might not focus on productivity in this context of learning, but rather emphasize the importance of growth.

One benefit of using GenAI in learning processes is the potential for providing immediate learning support, such as when using LLM-based tutors. However, concerns remain about the adaptability and integration of these tools into the learning process. Rigorous studies are needed to quantify growth, productivity, and other measures to determine whether AI is fulfilling its promises.

Strand 3. How does GenAI use affect personal and professional identity development?

In the long run, the use of GenAI may significantly impact identity development and *self-concept*. For instance, software engineers may experience an identity crisis by feeling threatened in their self-concept when their employers enforce the use of GenAI.

Widespread encouragement of GenAI use may instill diffuse and multifaceted fears among employees, encompassing concerns about job security, missing out or being left behind, and diminished self-value. Employees may worry about becoming dependent on GenAI or question whether failing to keep up with GenAI tools will make them inefficient and obsolete. Professionals who took pride in being selected for their unique skills may feel devalued when they see that GenAI can perform similar tasks. Thus, the use of GenAI can potentially downgrade one's sense of self-worth, challenging their professional identity and self-concept and leading to a sense of diminished self-value.

Pointing out ways of how to use and possibly when to avoid AI while preserving and cultivating one's identity, self-worth, and self-value within the profession could be crucial. Factors such as sense of belonging, self-efficacy, self-concept, and exploration of causes and comparators driving fears and anxieties may become more relevant in CS education than they are at this point in time.

Strand 4. How can GenAI be used ethically?

The use of GenAI presents several inherent ethical challenges; most pressingly, do we want to participate in a fundamentally unethical system? The value of using GenAI must be carefully considered for every context and application.

Our discussion also touched on the broader issue of ethical hypocrisy, such as the environmental impact of activities like flying and using ChatGPT. It was noted that ethics often involve making choices, and our responsibility as educators includes informing students and guiding them to understand the trade-offs involved in using GenAI so they can make

conscious decisions. Human behavior is often a result of negotiations with others, and education should support the comprehensive personal development of students. This involves identifying important values and teaching students how to uphold them, even in the face of disruptive technological advancements, such as GenAI tools.

Conclusions

Using GenAI in education may present both opportunities and challenges. Based on our discussion, we identified several areas of research that require further investigation. These include (i) gathering sound empirical evidence to evaluate the advantages and disadvantages of GenAI use and developing strategies to integrate GenAI into educational practices in a pedagogically meaningful way, while preserving the benefits of *learning struggle* and *critical thinking*; (ii) quantifying the impacts of GenAI use, developing guidelines for its integration, and exploring how GenAI learning processes can align with *human values* and *growth*; (iii) identifying and developing strategies to support individuals in maintaining a positive *self-concept* and fostering their *professional identity* in the face of technological change; and (iv) identifying educational and psychological mechanisms that encourage students to develop a more responsible and *value-driven approach* towards GenAI use.

References

1 Bjork, E. L. and Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, and J. R. Pomerantz (eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pp. 56–64. Worth Publishers. https://psycnet.apa.org/record/2011-19926-008

2 Hazzan, O. and Erez, Y. (2025). Rethinking computer science education in the age of GenAI. *ACM Transactions on Computing Education*, *25*(3). https://doi.org/10.1145/3732792

3 Kohn, T. (2025). From imitation games to robot-teachers: A review and discussion of the role of LLMs in computing education. *Journal of Computer Assisted Learning*, *41*(3), e70043. https://doi.org/10.1111/jcal.70043

4 Scholl, A. and Kiesler, N. (2024). How novice programmers use and experience ChatGPT when solving programming exercises in an introductory course. In *2024 IEEE Frontiers in Education Conference (FIE)*, Washington, DC, USA, pp. 1–9, IEEE. https://doi.org/10.1109/FIE61694.2024.10893442.

5 Zeißig, A., Kansok-Dusche, J., Fischer, S. M., Moeller, J., and Bilz, L. (2024). The association between boredom and creativity in educational contexts: A scoping review on research approaches and empirical findings. *Review of Education*, *12*(1), e3470. https://doi.org/10.1002/rev3.3470

4.3 We're at a Crossroads: How GenAI Presents Challenges to Equity and Inclusion in Computing Education

Earl Huff (University of Texas – Austin, US), Laura E. Brown (Michigan Technological University – Houghton, US), and Daniel Schiffner (DIPF – Frankfurt am Main, DE)

Advances in Generative AI (GenAI) have changed the way we view programming and programming education. GenAI's ability to write usable, working code has now sparked conversations and research regarding how computer science courses should be taught. Students

are utilizing tools such as ChatGPT to produce coded solutions to tasks, but still require the ability to read and comprehend the code. Current research examines the implications of GenAI tools across the breadth and depth of CS courses and what it means for future iterations of CS curricula. What is lacking, however, is research addressing GenAI's impact on future software developers on two fronts. On one front, we need to consider how differential access to GenAI tools and education in K-12 among specific populations can widen the gap in terms of academic progression in higher education, especially for historically marginalized groups. There is already a gap in access to curricula and teachers, as seen with the AP Computer Science Principles. GenAI may further widen this gap due to several factors: a lack of a curriculum adapted for GenAI, a shortage of teachers skilled in using, advising, and understanding GenAI, and limited access to computers and AI tools in schools. These factors can create disadvantageous situations for students entering college to pursue a degree in CS. On the other front, industry hiring practices are evolving unevenly, with AI being used both by candidates and evaluators, often without clear standards or awareness of the implications. To what extent may companies utilize GenAI tools to aid in the evaluation of candidates, and how will they try to mitigate bias in the hiring process? What are candidates allowed to delegate to an AI if applying for a position and what training do they require for it?

This position paper argues that without systemic changes to the entire education through hiring pathway, AI risks reinforcing existing biases, introducing new biases, and rendering traditional assessments obsolete. We call for a rethinking of programming education across all levels, so that students acquire the competencies to utilize AI, and avoid the creation of new biases and unethical systems by gatekeeping newcomers who do not have access to special tools.

Participants

- Ibrahim Albluwi
Princess Sumaya University for Technology – Amman, JO
- Imen Azaiz
LMU München, DE
- Jamie Benario
Google – Chicago, US
- Dennis Bouvier
United States Air Force Academy, US
- Claus Brabrand
IT University of Copenhagen, DK
- Laura E. Brown
Michigan Technological University – Houghton, US
- Neil Brown
King's College London, GB
- Michelle Craig
University of Toronto, CA
- Paul Denny
University of Auckland, NZ
- Rodrigo Duran
Federal Institute of Brasília, BR
- Carolin Hahnel
Ruhr-Universität Bochum, DE
- Earl Huff
University of Texas – Austin, US
- Christopher D. Hundhausen
Oregon State University – Corvallis, US
- Amanpreet Kapoor
University of Florida – Gainesville, US
- Hieke Keuning
Utrecht University, NL
- Natalie Kiesler
Technische Hochschule Nürnberg, DE
- Tobias Kohn
KIT – Karlsruher Institut für Technologie, DE
- Dennis Komm
ETH Zürich, CH
- Juho Leinonen
Aalto University, FI
- Colleen Lewis
University of Illinois Urbana-Champaign, US
- Kevin Lin
University of Washington – Seattle, US
- Dominic Lohr
Universität Erlangen-Nürnberg, DE
- Andrew James Luxton-Reilly
University of Auckland, NZ
- Stephen MacNeil
Temple University – Philadelphia, US
- Victor-Alexandru Padurean
MPI für Software Systems – Saarbrücken, DE
- Leo Porter
University of California – San Diego, US
- James Prather
Abilene Christian University, US
- Brent Reeves
Abilene Christian University, US
- Karen Reid
University of Toronto, CA
- Daniel Schiffner
DIPF – Frankfurt am Main, DE
- Jan Schneider
DIPF – Frankfurt am Main, DE
- Adish Singla
MPI-SWS – Saarbrücken, DE
- David H. Smith IV
Virginia Polytechnic Institute – Blacksburg, US
- Jacqueline Staub
Universität Trier, DE
- Sven Strickroth
LMU München, DE
- Claudia Szabo
University of Adelaide, AU
- Shubbhi Taneja
Worcester Polytechnic Institute, US
- Christina Weers
Goethe-Universität – Frankfurt am Main, DE
- Michel Wermelinger
The Open University – Milton Keynes, GB
- Titus Winters
Adobe – New York, US
- Daniel Zingaro
University of Toronto Mississauga, CA
- Jaromír Šavelka
Carnegie Mellon University – Pittsburgh, US

Report from Dagstuhl Seminar 25312

Building Privacy-Preserving Technologies of Societal Impact

Marina Blanton[*1] **and Liina Kamm**[*2]

1 **University at Buffalo – SUNY, US.** mblanton@buffalo.edu
2 **Cybernetica AS – Tartu, EE.** liina.kamm@cyber.ee

Abstract

This report describes the motivation, purpose, and scope of Dagstuhl Seminar 25312 "Building Privacy-Preserving Technologies of Societal Impact" as well as documents its program and outcomes. This inter-disciplinary seminar brought together computer science researchers and practitioners working on building privacy-enhancing technologies – most notably secure computation applications – and researchers in expertise in other relevant disciplines including law, medicine, and social studies. Besides the applied nature of the seminar that capitalized on the participants' desire to facilitate adoption of privacy-enhancing techniques in real world applications, a unique aspect of this seminar was the shared passion of the participants to use their expertise to build tools for protecting vulnerable populations and for other public good purposes.

Seminar July 27 – August 1, 2025 – https://www.dagstuhl.de/25312
2012 ACM Subject Classification Security and privacy → Cryptography; Security and privacy → Human and societal aspects of security and privacy; Applied computing → Law; Information systems → Data management systems
Keywords and phrases Privacy-enhancing technologies, applications, societal impact, secure computation
Digital Object Identifier 10.4230/DagRep.15.7.280

1 Executive Summary

Marina Blanton (University at Buffalo – SUNY, US, mblanton@buffalo.edu)
Liina Kamm (Cybernetica, EE, liina.kamm@cyber.ee)

The main goal of the seminar on Building Privacy-Preserving Technologies of Societal Impact was to bring together the main actors in the privacy enhancing technology (PET) sphere to share knowledge of the current state of the art in PETs, get an overview of the most recent cases where these technologies have been used to facilitate data analysis in sensitive social topics, and to discuss ways how PETs could further be exploited to create even more societal impact.

Privacy-preserving techniques such as secure multi-party computation and related areas have matured over the last decades in terms of their speed, accessibility, and usability. However, their propagation into everyday user products continues to be slow. PETs remain largely inaccessible to end users and small or non-profit organizations. A large unexplored potential remains.

Researchers in the community have applied PETs to many application domains and have demonstrated that it is possible to contribute to solving large global challenges such as fighting crime, advancing medical research and patient treatment, strengthening sustainability efforts, reducing gender and race disparities, and much more. In all of this, the fact that certain

[*] Editor / Organizer

information remains confidential is crucial to enabling the functionality which otherwise would not be feasible to carry out. Thus, applied cryptography enables us to achieve what we could not do before using conventional mechanisms and improve individuals' wellbeing.

This seminar's goal was to bring together researchers, whose technical work has contributed to addressing societal challenges, and their allies – and it achieved the goal. The seminar welcomed 28 researchers and practitioners with expertise in computer science, medicine, and law. In their presentations, the participants shared their experiences in building PETs that protect vulnerable populations and contribute to public good as well as their experience with real-world deployment of PETs. The topics included supporting investigative journalists, building privacy-friendly humanitarian aid distribution, privately assessing developmental delays in toddlers, privately matching organ donors, performing medical research, and much more. The presentations also discussed the participants' experience with working with domain experts to design the solutions and deploying these and other applications in the real world. The abstracts of these talks are included in this report.

The participants identified in a brainstorming session a number of open questions they would like to discuss as part of the seminar. The questions ranged from adoption challenges and future strategies for application adoption to non-cryptographer's misconceptions and guidelines for addressing them, to more technical aspects such as interfaces and an evaluation framework, and to legal and funding aspects. The questions were assigned to different discussion sessions, grouping related questions to facilitate coverage of as many topics as possible. The results of these discussions are included in this report. A brief summary is that we need to educate the end users in choosing the right PETs, to convince policy makers of the necessity of PETs, to seek sources of funding for PETs that solve socially sensitive problems, to have the community create and adopt a comparative PET framework for objectively evaluating and comparing different solutions, and to capitalize on successes of privacy applications deployed at a large scale.

This research meeting was very productive and engaging. It had the level of engagement that cannot be found in other settings. And most importantly, it was the passion of the attendees – with technical expertise but deep care for addressing societal challenges – that resulted in the gathering to be hugely successful.

2 Table of Contents

3 Overview of Talks

3.1 Privacy by Design: Supporting Investigative Journalists via PETs

Kasra EdalatNejad (TU Darmstadt, DE)

Investigative journalists collect large numbers of digital documents during their investigations. These documents can greatly benefit other journalists' work. However, many of these documents contain sensitive information. Hence, possessing such documents can endanger reporters, their stories, and their sources. Consequently, many documents are used only for single, local, investigations. We presented DatashareNetwork, a decentralized and privacy-preserving search system that enables journalists worldwide to find documents via a dedicated network of peers, as the first search engine designed by journalists for journalists in 2020 to address this problem. We start the talk by introducing real-world problems that investigative journalists face and describe DatashareNetwork as a possible solution. Then, we discuss the practical challenges of moving forward from an academic prototype to deploying DatashareNetwork for the International Consortium of Investigative (ICIJ). This talk covers (1) our joint requirement gathering and (2) design with journalists, (3) a user study to help ICIJ with presenting the privacy property of our system to journalists and making utility/privacy trade-off decisions, (4) deployment challenges to integrate DatashareNetwork into ICIJ's IT infrastructure, and finally (5) open problems that require more attention from the community.

3.2 Private Set Intersection for the Society

Thomas Schneider (TU Darmstadt, DE)

Joint work of Thomas Schneider, members of the ENCRYPTO group & collaborators

Private set intersection allows two parties to privately compute the intersection of their inputs or even variants of this functionality. The first PSI protocol is due to C. A. Meadows (IEEE S&P'86) based on Diffie-Hellman and since then a lot of fruitful research has happened in this area, making it more and more practical. In this talk, I will summarize some of our works that investigate applications of PSI of societal impact.

In the first part, based on Kales et al. (USENIX Security'19) and Hagen et al. (NDSS'21), we look at contact discovery, which is a feature in almost all mobile messenger apps. Here, the goal is to match a few 1000 address book contacts against potentially billions of users of that service to determine which of the contacts are using the service. We show that the approach of hashing phone numbers as used by many messengers provides literally no additional privacy because phone numbers are structured and have little entropy, so they can be inverted within milliseconds. Protocols for unbalanced PSI provide the required privacy guarantees, but still, communication is a major bottleneck unless resorting to the multi-server setting, as we showed in Hetz et al. (ESORICS'23). Moreover, we show that the contact discovery functionality as implemented in the messengers can be exploited to mass-collect highly sensitive information like profile pictures or online status for a large number of phone numbers, and after our responsible disclosure the providers adjusted their rate limits.

The second part is based on Treiber et al. (WPES'22). This is an interdisciplinary effort together with law experts from Goethe University Frankfurt and the Police Academy Hamburg, where we investigate how law enforcement agencies can selectively and privately exchange data from both a legal and a technical perspective. We propose a system for lawful information exchange between law enforcement agencies using MPC and private set intersection, which has much higher privacy guarantees than the currently discussed approaches like pooling data in the clear in a data warehouse or exchanging hashed identities.

3.3 Privacy-Preserving Humanitarian Aid Distribution

Wouter Lueks (CISPA – Saarbrücken, DE)

Humanitarian aid-distribution programs help bring physical goods to people in need. Traditional paper-based solutions to support aid distribution do not scale to large populations and are hard to secure. Existing digital solutions, on the other hand, risk creating serious harms to recipients by collecting large amount of personal information including sometimes biometric data.

This talk covers our collaboration with the International Committee of the Red Cross to design digital aid-distribution systems with strong protection. We'll see how we used privacy-enhancing technologies to prevent harm to vulnerable aid recipients; and touch on our experience working with humanitarian organizations.

3.4 Regional Risk Monitoring of Developmental Delay in Toddlers Using MPC: Use Case Overview, and Aspects of Getting Secret-Shared Inputs via the Browser

Niek Bouman (Roseman Labs – Utrecht, NL)

The first part of this talk describes a real-world use case of one of Roseman Labs' clients, the Municipality of Rotterdam (the Netherlands), about monitoring the effectiveness of municipality-level policy concerning developmental delay in toddlers. You will learn how MPC was employed to achieve GDPR-compliance and how its use shortened the feedback cycle from 1,5 year to one month; creating real impact because the toddlers directly benefit from this.

The second, more technical part is about getting input data for MPC (in the "outsourcing model") from users via a web application, and discusses the problem of ensuring integrity of the web app, e.g., in a scenario where the web server gets compromised.

3.5 Addressing the Kidney Exchange Problem – From Theory Towards Practice

Susanne Wetzel (Stevens Institute of Technology – Hoboken, US)

Today, thousands of patients in Germany alone suffer from severe kidney disease and are on the waiting list to receive a kidney transplant. Receiving a living kidney donation is an alternative to waiting for a post-mortom kidney transplant. The main challenge with living kidney donation, however, is to find organ donors who are medically compatible with the patients. Also, not all countries allow kidney exchange due to the fear of manipulation, corruption, and coercion. Yet, in many countries where kidney exchange is allowed, there already are centralized platforms to facilitate the exchange – which in turn raises security and privacy concerns.

Our work seeks to develop privacy preserving protocols for solving the kidney exchange problem aiming to address these concerns.

In this talk, we discussed the challenges we encountered in carrying out this international and interdisciplinary project.

This is joint work with the research group of Professor Ulrike Meyer at RWTH Aachen University.

3.6 Secure Multiparty Computation in a European Clinical Study

Hendrik Ballhausen (LMU – München, DE)

Main reference Hendrik Ballhausen, Stefanie Corradini, Claus Belka, Dan Bogdanov, Luca Boldrini, Francesco Bono, Christian Goelz, Guillaume Landry, Giulia Panza, Katia Parodi, Riivo Talviste, Huong Elena Tran, Maria Antonietta Gambacorta, Sebastian Marschner: "Privacy-friendly evaluation of patient data with secure multiparty computation in a European pilot study", npj Digit. Medicine, Vol. 7(1), 2024.
URL http://dx.doi.org/10.1038/S41746-024-01293-4

In multicentric studies, data sharing between institutions might negatively impact patient privacy or data security. An alternative is federated analysis by secure multiparty computation. This pilot study demonstrates an architecture and implementation addressing both technical challenges and legal difficulties in the particularly demanding setting of clinical research on cancer patients within the strict European regulation on patient privacy and data protection.

3.7 Screening DNA Synthesis Orders for Hazards: Efficiently and with Privacy

Carsten Baum (Technical University of Denmark – Lyngby, DK)

DNA synthesis has become a ubiquitous tool in biological and medical research. It allows to synthesize arbitrarily built DNA strands based on digital descriptions only. However, DNA synthesis can be used for nefarious purposes.

The SecureDNA project built an efficient software tool to screen DNA orders for hazardous sequences with some privacy guarantees. In this talk, I describe the general architecture, the development process and some cryptographic underpinnings.

3.8 The German Self-Determination Act: MPC as a Tool for Better Privacy in Administrative Data Exchange?

Andreas Brüggemann (TU Darmstadt, DE)

Joint work of Linda Seyda, Andreas Brüggemann, Gerrit Hornung, Thomas Schneider
Main reference Linda Seyda, Andreas Brüggemann, Gerrit Hornung, Thomas Schneider: "Multi-Party Computation als Instrument zur Umsetzung datenschutzkonformer behördlicher Datenabgleiche: Eine interdisziplinäre Analyse am Beispiel der Diskussionen um das Gesetz zur Selbstbestimmung über den Geschlechtseintrag", in Proc. of the 54. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2024 - Lock in or log out? Wie digitale Souveränität gelingt, Wiesbaden, Germany, September 24-26, 2024, LNI, Vol. P-352, pp. 153–167, Gesellschaft für Informatik, Bonn, 2024.
URL http://dx.doi.org/10.18420/INF2024_11

In 2024, the German self-determination act (SBGG) was passed, enabling transgender, intersex, and non-binary people to change their name and gender entries by self-declaration. The original draft of the SBGG included that all entry changes are automatically forwarded to many law enforcement agencies to allow them to keep their internal databases up to date. This data transfer would have been unprecedented with name changes due to other reasons not being transferred to law enforcement without cause. Furthermore, it would have amplified the risk of lists of transgender, intersex, and non-binary people being created in a situation where German police has already been reported to collect lists of queer people in the past while also, violence against queer people has recently been increasing.

In an interdisciplinary collaboration on the intersection of law and cryptography, we have analyzed the risks of the planned data transfer as well as the underlying considerations from the perspective of data protection. Not only is the proportionality of the transfer generally at least highly questionable, but the transfer also contains data about individuals completely unknown to law enforcement where a legitimate interest on the side of law enforcement cannot apply. Yet, cryptographic means such as private set intersection (PSI) have been ignored in the lawmaking process while they would have enabled to at least protect information about most affected people. We provide an analysis of how PSI could have improved parts of the problem, which problems it could not have resolved, and which smaller, but new risks it might have introduced. While the data transfer has eventually been removed before the law was passed, this was only done for consistency reasons with the plan to reintroduce a similar mechanism for all name changes in Germany while an even worse version is now being discussed specifically for the SBGG. Our research in this lawmaking process showcases how concrete risks and data protection concerns were ignored while generally, privacy enhancing technologies such as PSI appear not to be considered or even known to the responsible lawmakers while they would enable possible compromises.

3.9 Verifiable Carbon Accounting in Supply Chains

Jonathan Heiß (TU Berlin, DE)

Main reference Jonathan Heiss, Tahir Oegel, Mehran Shakeri, Stefan Tai: "Verifiable Carbon Accounting in Supply Chains", IEEE Trans. Serv. Comput., Vol. 17(4), pp. 1861–1874, 2024.
URL http://dx.doi.org/10.1109/TSC.2023.3332831

Trustworthy data sharing in carbon accounting is hindered by confidentiality constraints that prevent consumers from understanding how providers construct emission data. This opacity enables greenwashing and undermines consistency across value chains. Zero-knowledge proofs (SNARKs) combined with verifiable data structures, and other privacy-enhancing technologies (PETs) address these limitations by enabling verification without revealing sensitive business internals. This talk introduces verifiable carbon accounting (VCA) as a framework synthesizing multiple PETs for scalable, confidentiality-preserving emission data verification and discusses adoption strategies to integrate VCA into existing accounting standards as an alternative to costly, non-scalable third-party verification.

3.10 MPC in Practice: Lessons from Germany's Public Sector and Future Applications in Sustainability

Ágnes Kiss (SINE Foundation – Berlin, DE)

This presentation explores the practical implementation of secure Multi-Party Computation (MPC) through real-world use cases in Germany's public sector and emerging applications in sustainability.

Our work with German municipalities reveals that successful MPC deployment requires more than technical solutions. While trust between parties was not the primary barrier, data privacy concerns and GDPR compliance proved crucial. We present concrete use cases including disaster management, childhood health examination analytics, and vaccination verification, highlighting the importance of finding committed stakeholders willing to pioneer new approaches. The technical foundation relies on our MPC framework combining Garble (high-level Rust-like programming for Boolean circuits) and polytune (full-threshold MPC with authenticated garbling).

Looking toward sustainability applications, we identify significant potential for privacy-preserving technologies in carbon footprint verification and supply chain transparency. Product Carbon Footprint (PCF) calculations require verifiable accuracy and transparency, yet involve sensitive trade secrets across multiple stakeholders. MPC offers a path to shared insights from primary data without compromising competitive advantages, enabling e.g., statistical outlier detection, PCF categorization, and input validation while preserving data confidentiality. The potential of Privacy-Enhancing Technologies in sustainability contexts remains largely unexplored, presenting opportunities for interdisciplinary research and practical applications.

3.11 Secure Analytics with MPC Made Practical: Lessons Learned from Medical and Workforce Data Use Cases

John Liagouris (Boston University, US)

"The performance of MPC-based approaches is so low that practical applicability is not in sight." This is a review excerpt of a paper I co-authored, describing our vision to use multiparty computation (MPC) for secure data analytics in the cloud. In this talk, I will share how – 4+ years later – we have realized this unlikely vision and more. I will first explain the legitimate skepticism of the particular reviewer and why past results indicated that MPC protocols were impractical for complex analytics. I will then argue that careful system design and cross-layer optimizations can not only amortize MPC costs, but also achieve scalability to large inputs and complex workloads, without compromising security. I will present the BU Secure Analytics Stack, our unified software architecture for secure collaborative data analysis, and discuss some lessons learned from real use cases. Finally, I will show performance results for secure relational and time series analytics at a scale that a few years ago was only possible with information leakage or the use of trusted compute.

3.12 MPC for Securing Auctions in a P2P Electricity Market

Mariana Gama (KU Leuven, BE)

The development of the smart grid and introduction of smart metering devices facilitates the emergence of new applications within the energy domain, among which is the p2p electricity trading market. This market would allow owners of renewable energy sources to sell any excess electricity generated directly to other users, incentivising users to acquire renewable energy sources and diminishing transmission losses by promoting electricity exchanges among neighbours. However, there are several privacy concerns associated with the p2p electricity market, as exposing fine-grained metering data exposes sensitive information about users' habits.

This talk introduces MPC-based auction mechanisms for the future p2p electricity market. We discuss the possibility of prioritising both low-volume orders and close neighbours when matching orders for intraday p2p electricity trading, and propose a day-ahead flexibility market where users' consumption schedule is optimised with respect to day-ahead electricity prices.

3.13 PETs in Practice

Liina Kamm (Cybernetica AS – Tartu, EE)

Privacy enhancing technologies (PETs) have been around for decades. Their feasibility has been an area of research for tens of research teams. However the practical uptake of the more complex and privacy-preserving solutions has been slow.

In the beginning of 2023, Estonia conducted a research project on privacy enhancing technologies to work out a concept and roadmap for deploying these technologies in e-government. We interviewed people from 18 state agencies to find out their expectations and requirements for the use of different PETs. This lead us to compile a PET concept that gives and overview of the technologies and describes the generalised usage archetypes for e-government. The PET reoadmap gives a concrete way forward.

We discussed the lessons learned from this work and distilled from this a list of incentives and barriers that possible users and customers see.

3.14 Securing Satellite Rendezvous and Proximity Operations with MPC

Kevin Butler (University of Florida – Gainesville, US)

Space is emerging as a contested and congested environment, driven by the accelerating deployment of satellites from government and commercial organizations. Consequently, in-space security and privacy have become a significant concern, particularly related to satellite rendezvous and proximity operations (RPO). This talk describes our efforts to develop and implement critical RPO algorithms on radiation-tolerant hardware suitable for deployment in low Earth orbit using MPC to identify and ensure privacy of appropriate inputs, implemented with the MP-SPDZ framework. We highlight the need for optimizations and careful mission considerations when deploying MPC in this specialized environment. From a social impact perspective, we also discuss the potential for MPC to address interpersonal threats posed by continuous location-sharing services.

3.15 Scaling Privacy: Cloud-Native MPC, Data Anonymization, and the Path to Open Linux Ecosystem

Hossein Yalame (Robert Bosch GmbH – Renningen, DE)

As organizations increasingly rely on data-driven decision-making, preserving privacy at scale has become a critical challenge. This work presents CarbyneStack, a cloud-native framework for secure multi-party computation (MPC), developed and deployed at Bosch to enable privacy-preserving analytics across global subsidiaries. Bosch has contributed CarbyneStack

to the Linux Foundation Europe, fostering an open ecosystem for privacy-enhancing technologies. We discuss how this open-source initiative promotes interoperability, transparency, and community-driven innovation while maintaining enterprise-grade performance. Our results highlight that having MPC in a cloud-native architecture not only scales privacy-preserving analytics but also accelerates the adoption of secure, privacy-focused solutions across industries.

3.16 Increasing the Impact of an MPC (or Any PET) Application

Dan Bogdanov (Cybernetica AS – Tartu, EE)

What is needed to grow the impact of a Privacy Enhancing Technology system?

Easier to achieve:

- People with skills – PET and MPC skill dissemination works well
- Open code – we found 56 projects/products (not all open, but there are more)
- Funding for pilots – EU and US have both invested > 300M€/$ into PETs by now!
- Regulatory support – it is doable for non-PET solutions, so it's a matter of will

Harder to achieve:

- Sponsor/customer deciding to use PETs – fears of cost, alternative solutions
- Working code – keeping cryptography code running for years is non-trivial
- Ongoing operations – need to find money and skills to run the (distributed!) system
- Trust building – need to convince end users and the public that this will be secure

The talk focuses on progress in the hard-to-achieve aspects of PET growth.

3.17 Six Years of MP-SPDZ in the Wild

Marcel Keller (CSIRO – Eveleigh, AU)

In the few years since publication, MP-SPDZ has received more than 1000 GitHub issues and hundreds of citations. I will present a categorization of issues and citations that mention the usage of the framework. I will also enumerate the most frequent themes in the issues and give my view on how to address them, whether it's worth it, or possible at all. This is to gather views from the community and foster discussion.

3.18 Getting up to (MP-)SPDZ – Challenges Faced by Developers Getting into MPC

Vincent Ehrmanntraut (RWTH Aachen, DE)

For most people, MPC is a dauntig subject to approach at first, as a master of algorithmic concepts and implementation details is needed to design effective protocols. This results in a very steep learning curve.

In this talk, I share insights from advising bachelor and master theses to point out specific pain points and present some ideas that might help to flatten the learing curve. The talk then pivots to make a case that the whole field of (high-level) MPC protocols needs to establish evaluation practices that improve the comparability of protocols across papers.

3.19 Challenges of Making PETs with Societal Impact Usable & Trustworthy

Simone Fischer-Hübner (Karlstad University, SE)

Privacy Enhancing Technologies (PETs) will only be successfully deployed if they are usable and trustworthy. However, for making PETs usable, various challenges need to be addressed: First of all, there are no "one size fits all" solutions. The context of a data processing applications and demographics of users, and in particular their cultural and gender backgrounds, need to be considered for the usable design of PETs. Secondly, explaining PETs that are based on "crypto magic" operations constitutes a challenge and additionally technical background knowledge may have a negative impact on the users' mental models. Thirdly, for the configuration of PETs, interdisciplinary expertise may be needed and complex trade-offs between protection goals may have to be made. In this talk, these usability challenges are illustrated with six user studies on PETs that can have societal impact, including a Selective Authentic EHR Exchange Service based on malleable signatures, privacy preserving data analytics based on homomorphic encryption, differential privacy and functional encryption, and lastly secret sharing for secure cloud storage. The presentation concludes with guidelines and recommendations for usable and trustworthy PETs.

3.20 How People would Trust a Government based on its Choice of a Digital Identity Wallet?

Kazue Sako (Waseda University – Tokyo, JP)

I would like to share preliminary results from a survey we conducted in five countries, each with 800 participants, exploring how people would trust a hypothetical government based on its choice of a digital identity wallet – especially when the wallet supports unlinkable selective disclosure. The study was conducted with a social psychology researcher, who carefully designed the survey and handled the statistical analysis.

3.21 MPC and PETs from a GDPR Perspective

Meiko Jensen (Karlstad University, SE)

Multiparty Computations are commonly perceived as a privacy-enhancing technology of high quality. However, when it comes to compliance to data protection laws like the European GDPR, the implementation of MPC schemes may have unintended side effects. In this talk we discussed different aspects of the GDPR and their relation to PETs in general and MPC specifically. Here, we considered three cases: MPC as an anonymization tool, MPC as a risk reduction tool according to Art. 35 GDPR, and MPC as a methodology to implement Data Protection by Design (Art. 25) and Security of Processing (Art. 32).

For anonymization, we noted that the intermediate results of an MPC computation may still contain linkable information of a data subject, hence, for most cases, applying MPC does not result in a sufficient level of anonymization as defined in the GDPR. Hence, MPC should not be considered as an anonymization tool in order to get out of the scope of the GDPR.

With respect to risk reduction, we noticed that the implementation of MPC does in fact help with confidentiality and integrity risk mitigation, but its implementation opens up several other risk vectors, such as availability issues (due to performance and network workloads), administration issues, and issues related to the addition of new stakeholders (the compute nodes that may belong to different organizations, hence causing either joint controllership or controller-processor relationships in the sense of GDPR). Hence, applying MPC is a trade-off of risk mitigation and novel risk addition, which should be evaluated carefully.

Finally, though MPC may clearly be seen as a sound approach to implement privacy by design and security of processing, it nevertheless is not mandatory to be implemented. Other PETs of lower quality may suffice to cover these GDPR requirements, so the incentive to utilize MPC, given its novel risk additions and pitfalls, must be evaluated carefully.

Based on these observations, we discussed different approaches to improve the status of MPC under GDPR considerations. We identified that it needs more success stories of MPC implementation, to get out of the "PET graveyard" of piloted, but never marketed MPC tools. Also, an update to GDPR may help catering for the specific needs of MPC adoption. Finally, we identified that MPC is one of the very few privacy-enhancing technologies that is capable of protecting against the "trump attacker model," where a government turns rogue against its own institutions and tries to misuse existing data, including the removal or bypass of protection mechanisms (e.g., by forcing decryption and secret key revelation). If one or more parties of the MPC implementation are out of scope of such an attacker, e.g., in a different jurisdiction, MPC could withstand such an attack incident.

3.22 Coordinating MPC Parties with a Byzantine Fault-Safe Distributed Database

Mark Abspoel (Roseman Labs – Utrecht, NL)

Consider an outsourced secure multiparty computation scenario with a small number of long-running computation parties that can execute multiple computations, as follows. Input parties asynchronously secret-share confidential data and distribute the shares to the computation parties. An output party can request for a computation to be run on the secret-shared input data, subject to manual approval by a fixed set of approver users.

This talk is about how a distributed database of computation-related metadata can be facilitated by the computation parties, that may contain the schemas of secret-shared data, the approvals that are given by the approver users, audit logs of computations, et cetera. The database should be resilient against an active adversary corrupting a dishonest majority of computation parties.

We develop a novel Byzantine-fault-safe distributed database through a synchronizing layer on top of existing relational databases, that handles access control and ensures determinism. It is based on a novel protocol for ledger consensus with abort, where we circumvent the impossibility result for ledger consensus with a dishonest majority by replacing liveness by a weaker notion of liveness with abort.

3.23 Privacy-Preserving Collaborative Learning

Sinem Sav (Bilkent University – Ankara, TR)

In this talk, I'll share my experience through the evolving space of privacy-preserving collaborative learning. I'll reflect on the challenges we've faced in implementing the first federated and privacy-preserving machine learning framework and its application to the biomedical domain such as: (1) Conflicting application needs, (2) Inconsistent notions of privacy across domains, and (3) The growing tension between theory and practice.

Key discussion points will include: (1) Whether tailored, application-specific solutions are more effective than general-purpose ones; (2) How we might reconcile varying interpretations of "privacy-preservation" across stakeholders; and (3) What it will take to make techniques like homomorphic encryption or differential privacy more practical. I'll also mention the details that often get overlooked, like how to set up a collaborative learning pipeline, e.g., from hyperparameter tuning or data normalization.

3.24 They Drop, We Chat: Steganographic Censorship Circumvention for Chat-based Applications

Boya Wang (EPFL – Lausanne, CH)

Censorship prevents communications that are against the interest of the censor, and hence, silences the conversations key to a healthy society. Chat applications become a major target of censorship as they become the most popular tools for our communication nowadays. One type of censorship is sensitive-word filtering (SWF) which exists in popular non-E2EE chat applications such as WeiXin/WeChat. The censor deploys an adaptive blocklist of words at the application server to drop matched messages in real-time. In this work, we investigate steganographic circumvention system of content-based censorship in chat applications. Previous work studies the security of steganographic circumvention systems in isolation without considering system-wise or contextual requirements, which results for a mismatch between theoretical security guarantee and practical undetectability. We fill this gap by proposing a new design of the circumvention system, provide two steganographic instantiations, and conduct an empirical evaluation.

3.25 Understanding Differential Privacy Adoption Challenges for Technical Implementers

Rebecca Wright (Barnard College, Columbia University – New York, US)

Joint work of Liudas Panavas, Saeyoung Rho, Hari Bhimaraju, Wynee Pintado, Rebecca Wright, Rachel Cummings

Introduced by Dwork et al. in 2006, differential privacy can provide rigorous mathematical guarantees of user privacy. Differential privacy has been a research success and has been adopted by a number of large organizations, including companies like Apple, Google, and Microsoft, as well as the US Census Bureau and the Israeli Ministry of Health. Nonetheless, there are socio-technical challenges that have limited broad adoption.

In this talk, we describe our interview-based research study seeking specifically to address the question: For technical individuals newly introduced to differential privacy, which concepts are most challenging to grasp, and what factors contribute to these challenges? We interviewed 10 subjects, people knowledgeable about DP who have managed or worked with software engineers not previously knowledgeable about DP to implement it.

We discuss the four primary themes we find in the interviews: (1) scoping as a structural barrier, (2) the need for expert judgment, (3) tunable vs. codified parameters: as simple as possible, but not oversimplified, and (4) visualizations facilitate communication. We conclude with some behind-the-scenes stories of our project as well as possible next steps.

4 Discussion Sessions

4.1 Where Does MPC Have an Actual Advantage?

The topic of this discussion session was to determine where secure multi-party computation has advantage compared to other technologies and thus where it can see adoption success.

We discussed the following question: Assume trusted execution environments (TEEs) are secure, and that fully homomorphic encryption (FHE) is fast. Is there still a point to secure multiparty computation (MPC)?

We came up with three main advantages of MPC.

First, since MPC is software-based, anyone can inspect the source code, or even build an implementation based on a specification. This is in contrast to hardware-based solutions such as TEEs or FHE accelerators, where only some countries or organizations may have the capabilities to produce and/or inspect them. This point can be important due to geopolitical considerations, e.g., where there are different countries that need a symmetry of power.

Second, due to its distributed nature, MPC offers more control over which computations are executed on confidential data. For example, parties that supply data can choose to take on a role as a computation party, which enforces the need for its consent and cooperation (given a suitable security model).

Third, it is easier to fix vulnerabilities in MPC due to its software-based nature, which is especially important for long-running systems. By contrast, fixing vulnerabilities in hardware-based solutions can be costly and hard to achieve.

We also discussed a few aspects that hinder adoption of MPC. For lawyers, TEEs and MPC are relatively indistinguishable. This also applies to policy makers: because MPC is hard to understand, it does not really make its way into policies and regulations. For cloud providers, promoting MPC is risky, because it may lead users to infer that the cloud is not as secure as they claim. Consultancy companies may be aware of MPC, but it can also be difficult to come up with concrete use cases that generate business value.

As a side remark, it is interesting that MPC still has a place in key management, or is used as an alternative for a hardware security module, for applications with high security needs (e.g., cryptocurrency wallets). Perhaps some users do think MPC is more secure than hardware-based solutions.

4.2 How Should We Speak to Non-Cryptographers and How Can We Convince Medical Data Guardians?

Below is the summary of the discussion on guidelines on talking to non-crypto persons and developing a roadmap to talk to medical data guardians (who, in our experience, resist the use of MPC).

The most prominent theme of the discussion was that talking to non-crypto persons requires knowledge of the non-crypto domain, and a good understanding of the non-crypto person's pain points / problems.

Furthermore, demos in the context of the non-crypto person and general success stories, where MPC already is used successfully, are helpful to convince the non-crypto person that using MPC might be a good idea. On the other hand, telling the non-crypto person that there is a privacy problem, and then attempting to sell the non-crypto person on that privacy problem usually fails. The main exception is when the non-crypto person is already passionate about that privacy problem.

On a conversational level, non-crypto persons generally do not like the word attacker, or being called untrustworthy. A possible remedy is to introduce the "Trump Attacker," i.e., arguing that the successor of the non-crypto person might not be trustworthy.

Simplification of cryptographic concepts is another important aspect. Non-crypto persons usually need harder simplifications than initially assumed by cryptographers. Also, cryptographers tend to overly focus on the negatives of their technologies, e.g., by starting with the assumptions their systems require or directly pointing out possible attacks or implementation flaws. While it is important to remain realistic and not to oversell MPC, it may be necessary to initially mask uncertainty and have (faux) confidence in proposed solutions.

The discussion on healthcare-specific topics was very brief, and identified two problems: First, the medical persons might not see the need to go beyond legal compliance, e.g., with HIPPA. Secondly, hospitals (especially in the EU) already struggle with IT, and thus need extra convincing that adopting MPC won't burden their IT too much. Therefore, it might be necessary to "bring your own IT staff" for pilot projects.

4.3 What Misconceptions about Privacy-Enhancing Technologies Do Users Have?

We talked about what misconceptions users typically have about privacy technologies, and how to address some of them.

Many users are not able to distinguish well between different security technologies. For example, users may believe that privacy-enhancing technologies (PETs) like secure multiparty computation work akin more familiar cryptographic tools such as encryption. Or they think that end-to-end encryption also protects metadata. We have also seen that some users think that "private browsing" is a tool for connection security.

To address these misunderstandings, it is helpful to explain first what privacy guarantees are provided by a system, and to leave details about the privacy technology as secondary information.

For the privacy guarantees, we could come up with a more standardized way to, for a given system, communicate what data is accessible to what party, and under what guarantees (e.g., for WhatsApp, message content is accessible to a user's phone and not to the server, due to end-to-end encryption; but metadata is accessible to both the user's phone and the server). One question is who will write these: are creators of the systems going to do this, at the risk of presenting their system in a less favorable way, or is this best left to independent parties such as consumer organizations?

To explain PETs, we can make the distinction between input and output privacy [1], where input privacy is what could be guaranteed by a trusted third party and is generally easier to explain, and output privacy concerns statistical disclosure control and is harder to make precise. Different media, such as video explainers or games that illustrate a toy protocol, can also help, since people have different optimal learning methods. Good real-world analogies also work, but the potential pitfall is that people may take the analogies too far. Another pitfall is that any concept or explainer of good security can also be misappropriated to explain bad security, making it more difficult for users to distinguish between the two.

References

1 United Nations, *The United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*, `https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf`, 2023.

4.4 What are Properties of PET Projects of Social Impact?

Below is a summary of the discussion about properties of projects that have had social impact. Knowing such properties can be useful for building impactful applications in the future.

The discussion identified several notable PET success stories, ranging from end-to-end encryption in messaging apps to differential privacy implementations. The most successful cases share key characteristics: they provide seamless user experiences with zero additional effort required, achieve broad adoption through network effects, and address concrete adversary models. Examples like Signal's encrypted messaging, TOR for censorship avoidance, ad blockers, and Apple's commercial differential privacy rollout demonstrate how privacy technologies can gain critical mass when they either don't increase user difficulty or enable previously impossible functionality. The discussion differentiated privacy-focused technologies from security-focused ones such as HTTPS, Let's Encrypt, and certificate transparency, which were also were noted as successes that improved the overall privacy landscape through widespread deployment.

The group also examined what distinguishes successful from unsuccessful privacy technologies, highlighting that complexity and user friction are major barriers to adoption. Failed or limited examples like Mastodon's complicated setup, encrypted email's usability challenges, and electronic voting systems illustrate how even technically sound solutions can struggle without considering user experience and deployment incentives. The discussion raised important questions about whether imperfect privacy solutions are preferable to no solutions at all, noting the risk of "privacy-washing" for compliance purposes and the challenge that once a technology is deployed, better alternatives face significant adoption hurdles. Corporate interest, financial backing, peer pressure effects, and persistence over time emerged as critical factors, alongside the recognition that some privacy technologies may require accepting trade-offs between privacy, functionality, and adoption potential. The discussion concluded by considering the questions raised by the "Heilmeier catechism," used by the US DARPA agency as a means of assessing the potential of a research project, and how these questions are applicable to developing impactful research.

4.5 Who Should be Financing Public-Good Implementations (and Deployments)? Who is Going to Build, Run, and Maintain PETs Software?

This discussion session combined two questions and we start by addressing the question of who should be financing public-good implementations (& deployments) of PETs.

Before we can proceed, we need to ask ourselves what we exactly mean by "public good." By public good, we mean:

- there is a public benefit (as opposed to a situation where only a single individual, a small group of individuals, or ca ompany benefits from it);
- it addresses a societal problem;
- the PET enables a greater public good, and *creates new value*, and the latter is the main driver for adoption (not just "improved privacy").

With respect to public-good applications, we should view PET software or infrastructure as a public utility, like water and electricity, thus part of the (government-funded) basic infrastructure. We then tried to identify and list possible sources of funding for PET implementations/deployments:

- government (providing continued/sustainable public infrastructure funding);
- (industry) associations (;e.g., Stifterverband in Germany, or G-BA);
- specially created legal vehicles for collaboration (e.g., TMNL in the Dutch banking sector);
- foundations (e.g., SINE);
- (social impact) investors;
- government / EU as a (launching) customer;
- government / EU in the form of research funding, at various TRLs; (continuation is often a problem – a project typically goes to the "PET graveyard" after funding stops);
- crowd-funding / donations;
- multi-national collaborations (e.g., EU-US cyber intel sharing, Eurostat, United Nations).

We also identified several factors that can stimulate the adoption of PETs:

- "spreading the word" about the existence of PETs;
- education about the benefits of PETs;
- raising awareness, for example, through "privacy labels" (like food labels);
- regulation (like GDPR) and its enforcement;
- value creation;
- pressure from society (example: COVID contact tracing);
- cyber threats (international/cross-border).

To answer the question of who is going to build, run, and maintain PET software, we established a list of potential entities:

- foundations (like SINE);
- SME companies (like Cybernetica, Roseman Labs)
- large corporations (like Apple, Google, NTT, Alibaba);
- joint legal vehicles created by stakeholders for collaboration;
- cooperatives;
- public sector (e.g., Data Protection Agencies);
- universities;
- partly government-owned institutes (like Fraunhofer [DE], Alexandra Institute [DK], TNO [NL])
- associations (like WFA for AdTech).

4.6 When Do PETs Work and When Do They Not Work? When Can PETs be Privacy Washing Instead?

Good solutions are not easy to understand, and easy solutions are not good.

4.6.1 When Do PETs Work and When Do They Not Work?

MPC tends to work if you need several partners in your computation. In this scenario, the parties avoid having a single central database that already has all data, but are able to answer queries as if there were one. MPC is useful if you have multiple data holders and everyone agrees that data should not be kept in a single super-database, but there is also a need to compute on that data. However, the parties may not be willing (or allowed) to even secret share their data. Whether MPC would still work if a government changes and the new government is not privacy-friendly, depends on the setup. The number of parties is relevant (for instance, n in the range of 2–5, $n > 10$, contact tracing). In federated learning: the more parties, the better.

MPC may not be applicable in the supplier and business partner scenarios. Compute at one party, verify at n parties is challenging to realise in MPC; in this case, other tech might be more suitable.

There is a difference in tailor-made MPC solutions and general-purpose MPC solutions. Tailor-made MPC is time consuming to implement, can be error prone. However, it can also outperform general-purpose MPC.

In MPC, there can be cases where the participating parties are not equal; e.g., one has data, the other has computational power. Then the possibility of using MPC depends on the trust model: do the input parties trust the computing parties and the whole process?

MPC does not help with other security issues (e.g., input authenticity/integrity). In fact, it might hinder data integrity checks if special measures are not used.

If the government has seen all the data anyway (e.g., in the case of China currently), what is their motivation to utilise MPC or other PETs?

In some societies and countries there is the mentality of "I have nothing to hide" and they expect that this applies to everyone. However, you are not going to reach some of the target groups if you do not use PETs.

Personal data protection and the industry setting have different incentives (push for innovation). There would probably be value in banks sharing their data as even their aggregate information might be compromised. Hence, banks are cautious and might be willing to try out different PETs.

In federated learning one can see more tangible improvement: it is possible to learn on confidential data that was not accessible before. However, classical federated learning is not a PET as it does not protect privacy. If aggregation is secure, it might be considered a PET, but models still leak information. Encrypting gradients does not provide adequate protection, but the combination of homomorphic encryption and differential privacy might be successful.

4.6.2 When Can PETs Become Privacy Washing and the Results be Misused?

Federated learning has a privacy washing feel as you still take all the data in.

MPC for some functionality can also be privacy washing, as it depends what the end result will be used for and what it contains. Also MPC where the users have no control and the computing party decides everything (e.g., what to compute, what to publish). The privacy claim only holds when the computed functionality is restricted and the computing parties obey the boundaries of the allowed computations and the amount of leakage.

Differential privacy without disclosing the parameter epsilon or when epsilon is too big is definitely privacy washing. The central question is whether epsilon is used responsibly.

The elephant in the room is anonymisation to get out of the restrictions of GDPR. Often the methods used are not sufficient, the data is re-identifiable, and data donors' privacy is compromised.

Often synthetic data is generated simply by shuffling IDs. This does not work and can be reversed, but gives organisations the ability to say that they are using synthetic data. This is definitely privacy washing.

4.7 What Would a Comparative Evaluation Framework Look Like for MPC?

Here is the summary from the discussion towards establishing a comparative evaluation framework. The rationale behind this topic was that multiple research areas have a standard set of benchmarks (e.g., datasets or test cases) for evaluation of new developments in that area and a more systematic way of evaluating MPC constructions or tools would be valuable to have.

First, we identified multiple issues we commonly encountered in current protocol evaluations: Parameters, such as network settings, computing power, single vs. multithreaded implementations, etc., often lack clarity. Additionally, there is often ambiguity regarding what is actually measured, e.g., whether runtimes include the offline phase. We shared experiences on papers whose artifacts either contradicted details mentioned in the paper or where significant parts of the evaluation were missing.

As MPC is a very versatile field and covers many different application settings, strict rules that enforce benchmarks in consistent settings will not be adopted. Therefore, guidelines are needed. Authors should aim to adhere closely to the guidelines, and explicitly state and justify any deviations. This is inspired by successful applications of such guidelines in other fields, for example the NeurIPS reproducibility checklist or "datasheets for datasets."

Encouraging authors to be more explicit was a general theme, especially about input parameters and the nature of the actual computations as discussed above. The inclusion of metrics beyond the runtime, such as the number of communication rounds, the total data sent during computation, and, if applicable, solution quality (e.g., accuracy in PPML) should be encouraged. To facilitate this process, providing a common LaTeX template covering essential details and a checklist could be beneficial.

Note that this mostly focuses on the reporting of evaluation parameters, and not on the selection of the evaluation parameters. This emphasis stems from the necessity of establishing reporting standards first. However, there was also support for desk-rejecting MPC papers that evaluate on a single machine without artificial network restrictions.

Regarding guidelines for artifacts, there was a brief discussion on whether just having "high-level code" suffices, or whether MPC engines should also be included. The latter facilitates re-evaluations but might not always be possible, e.g., in the case of proprietary implementations like Sharemind. Nonetheless, publishing "high-level code" should already contribute to reproducibility and makes catching creative benchmarks easier.

Additionally, artifacts should include runtime logs with metadata, such as the network setting (and other parameters), for each data point. This information is crucial to maintain usability when archiving code and data (as mandated in Sweden), and also is useful to independently find answers to questions not discussed in the paper, or to clarify on online and offline benchmark times.

For some sub-fields of MPC, there might be more concrete benchmarking guidelines. For example, PPML papers should always report the accuracy, which needs to be measured on some standard datasets. Private Set Intersection also might allow for more standardized symmetric and asymmetric set sizes. Furthermore, the SPHERE project might be used to both provide computation platforms, and to make existing artifacts accessible. (It did not come up during the discussion, but Dagstuhl also has an artifact publishing platform called DARTS.)

Finally, we also discussed the guideline creation process. An organizer will need to coordinate an initial draft that undergoes iterative validation by lots of authors and reviewers before publication as a paper. Periodic revision of the guidelines every few years will probably be necessary. To promote the use of the guidelines, reviewers could initially encourage authors to implement the guidelines at venues with revision option, such as PETs.

There also was a brief discussion on the chance of pilot projects currently buried on the PET graveyard coming back from the dead when the time is right. For example, this was seen in both ML and MPC priorly. However, it is questionable whether protocols today would still be applicable in the future.

Participants

- Mark Abspoel
Roseman Labs – Utrecht, NL
- Hendrik Ballhausen
LMU – München, DE
- Carsten Baum
Technical University of Denmark – Lyngby, DK
- Marina Blanton
University at Buffalo – SUNY, US
- Dan Bogdanov
Cybernetica AS – Tartu, EE
- Niek Bouman
Roseman Labs – Utrecht, NL
- Andreas Brüggemann
TU Darmstadt, DE
- Kevin Butler
University of Florida – Gainesville, US
- Kasra EdalatNejad
TU Darmstadt, DE
- Vincent Ehrmanntraut
RWTH Aachen, DE
- Simone Fischer-Hübner
Karlstad University, SE
- Mariana Gama
KU Leuven, BE
- Jonathan Heiß
TU Berlin, DE
- Meiko Jensen
Karlstad University, SE
- Liina Kamm
Cybernetica AS – Tartu, EE
- Marcel Keller
CSIRO – Eveleigh, AU
- Ryo Kikuchi
NTT – Tokyo, JP
- Ágnes Kiss
SINE Foundation – Berlin, DE
- John Liagouris
Boston University, US
- Wouter Lueks
CISPA – Saarbrücken, DE
- Kajetan Maliszewski
Technische Universität Berlin, DE
- Kazue Sako
Waseda University – Tokyo, JP
- Sinem Sav
Bilkent University – Ankara, TR
- Thomas Schneider
TU Darmstadt, DE
- Boya Wang
EPFL – Lausanne, CH
- Susanne Wetzel
Stevens Institute of Technology – Hoboken, US
- Rebecca Wright
Barnard College, Columbia University – New York, US
- Hossein Yalame
Robert Bosch GmbH – Renningen, DE