CAL 2026

Cognition and Artificial Life



Třešť May 25th - May 27th

PROCEEDINGS OF THE 24TH CONFERENCE

Editors:

Juraj Hvorecký Karina Zamrazilová Kristína Malinovská Kryštof Havlík

Juraj Hvorecký
Karina Zamrazilová
Kristína Malinovská
Kryštof Havlík

editors

Kognice a umělý život
Cognition and Artificial Life
KUŽ / CAL 2026

2026
Czech Technical University in Prague

Kognice a umělý život – Cognition and Artificial Life – KUŽ / CAL 2026
(peer-reviewed conference proceedings)



Reviewers:

Ing. Vojtěch Brejtr
prof. Ing. Igor Farkaš, Dr.
Mgr. Nicholas Hananeia
doc. Mgr. Matěj Hoffmann, Ph.D.
Mgr. Juraj Hvorecký, PhD.
Mgr. Bc. Juraj Jonáš, Ph.D.
Mgr. Ľudmila Lacková Bennett, Ph.D. et Ph.D.
Ing. Miroslav Macík, Ph.D.
RNDr. Kristína Malinovská, PhD.
Mgr. Bc. Michal Sedlák, Ph.D.
Ing. Zdeněk Straka, Ph.D.
Mgr. Aleš Svoboda, Ph.D.
doc. RNDr. Martin Takáč, PhD.
Mgr. Matúš Tomko, PhD.
Mgr. Michal Vavrečka, Ph.D.
Prof. RNDr. Jiří Wiedermann, DrSc

Published in Prague, 2026

Publisher: Czech Technical University in Prague
Prepared by: Czech Institute of Informatics, Robotics and Cybernetics (CIIRC), CTU in Prague
Jugoslávských partyzánů 1580/3
160 00 Praha
Tel.: +420 224 354 224

Edited by: Mgr. Juraj Hvorecký, PhD.
Mgr. Karina Zamrazilová
RNDr. Kristína Malinovská, PhD.
Bc. Kryštof Havlík

Copyright: article authors

ISBN 978-80-01-07572-2
DOI: https://kognitivniveda.cz/cal-2026/
This publication is licensed under Creative Commons license
CC BY-NC-ND

Foreword

Dear colleagues and friends of cognitive science,

It is our great pleasure to present the proceedings of the 24th Conference on Cognition and Artificial Life 2026. We are delighted to continue the long-standing tradition of this established meeting of researchers, students, and enthusiasts in cognitive science, which was held for the third time in the inspiring setting of the Třešť Chateau.

This year's conference also provided an opportunity to commemorate the fifth anniversary of the passing of the distinguished Czech cognitive scientist Associate Professor Ivan M. Havel, whose scientific achievements and personal legacy continue to shape the development of cognitive science in the Czech Republic. We are deeply grateful that our invitation was accepted by our distinguished guests, Professor Václav Hlaváč and Dr. Dagmar Havlová. The first day of the conference therefore concluded with a commemorative session dedicated not only to the life and work of Associate Professor Havel, but also to the development of Czech cognitive science from its beginnings to the present day.

This year's conference is particularly remarkable because, for the first time in its history, it was organized predominantly by students and early-career researchers, representing the third generation of the Czech cognitive science community. We regard this as both a testament to the continuity of our field and a commitment to fostering interdisciplinary dialogue, openness, international collaboration, and generational renewal. At the same time, we continue the process of internationalization initiated during the previous edition of the conference in Slovakia. Consequently, the official languages of the conference were Czech, Slovak, and English.

The conference program covered a broad spectrum of topics across cognitive science, ranging from artificial intelligence, robotics, and computational modeling to linguistics, philosophy of mind, psychology, neuroscience, biology, and the arts. This thematic diversity reflects the very nature of cognitive science as an interdisciplinary field that brings together multiple scientific perspectives in the pursuit of a deeper understanding of the mind, cognition, and intelligent behavior.

All contributions included in these proceedings, whether presented as oral presentations or posters, underwent a peer-review process. We would like to express our sincere gratitude to all sixteen reviewers for their time, expertise, and thoughtful evaluations, which substantially contributed to the quality of this volume.

Finally, we would like to thank all authors, speakers, participants, partners, and members of the organizing committee whose dedication and enthusiasm made this conference possible. We hope that these proceedings will serve not only as a record of the ideas and discussions shared during the conference, but also as a source of inspiration for future research, new collaborations, and the continued advancement of cognitive science in the Czech Republic and beyond.

We wish you an inspiring and enjoyable reading experience.

Mgr. Juraj Hvorecký, PhD and Mgr. Karina Zamrazilová

Contents

Invited speakers 8

Modeling and Exploring Visual Concepts in Deep Networks
Mhd Walid Al Jallad, Kristína Malinovská 12

Auditory distance perception for sounds around the listener
Gabriela Andrejkova, Norbert Kopčo 17

Semantic Robustness of Unsupervised Approaches for Interpretability of Vision Language Models
Tamara Bílá, Adam Červenka, Igor Farkaš 21

Zkreslení pozornosti při percepci jídla: Eye-trackingová studie obézních dětí
Jiří Čeněk, Stanislav Mokrý, Lenka Hromková, Beatrice Elena Chromková Manea, Leona Mužíková, Veronika Suchodolová, Anna Svozilová, Dana Hübelová 26

Sex s roboty: perverze, parafilie a etika umělé intimity
David Černý 29

Constructing Grounded Conceptual Representations via Object-Centric Active Vision
Miroslav Cibula, Kristína Malinovská, Igor Farkaš 34

Společnost pro kognitivní vědu: komunita jako podmínka rozvoje kognitivní vědy v českém prostředí
Frederik Ďalak 39

Would Psychedelics Alter the Fitness Landscape of Cooperation? An Experimental Framework for an Agent-Based Model
Aly Debevec-Kruse, Martin Takáč 41

Associative Memory with Modern Hopfield Networks for Incremental and Open-Set Visual Learning
Jelena Epifanic, Kristína Malinovská 46

Unsupervised Topographic Representations of Motor Primitives in Humanoid Manipulation and Gestures
Radovan Gregor, Igor Farkaš 52

Multi-Scale Simulation of Synaptic Plasticity Induced by Repetitive Magnetic Stimulation
Nicholas Hananeia 56

Virtual reality enabling cognitive map acquisition for persons with visual impairment: A Systematic Literature Review
Kryštof Havlík, Miroslav Macík 61

AI-Enabled Big-Data Developmental Science: Longitudinal Analysis of Infant Sensorimotor Development
Matěj Hoffmann 69

Psychologické aspekty vývoja konverzačných agentov
Juraj Jonáš 73

Predicting Item Difficulty in German Language Tests: A Machine Learning Approach with Linguistic Features
Ivana Kapounová, Jana Dlouhá 78

Explorations in Predictive Coding Models of Hallucinations
Milica Kiš, Martin Takáč 85

Izhikevich Model of a Spiking Neuron: A Qualitative Analysis Through the Phase Plane
Hana Kollárová 90

Dynamický Turingův test: Adaptace rozhodovacıch strategií při detekci impresionistických maleb generovaných umělou inteligencí
Nikita Kostka, Ondřej Vadinský 97

Exploring the Tolman-Eichenbaum Machine for Modeling Spatial Understanding in Robots
Michal Kostrhun, Igor Farkaš 104

Analýza citlivosti genetických algoritmů na nastavení parametrů srovnání fixního a adaptivního přístupu
Anna Krajčířová 108

Does a transformer think like a human?
Andrej Lúčny 116

Falzifikační charakter sociální agence v užívání vizuálně-generativní AI
Jan Šmejkal 120

Mind Wandering during Lectures
Magda Stanová 124

Konceptuální redukce a smrt obrazu
Aleš Svoboda 129

Lidé vs. Algoritmy: Preferujeme stále ještě lidské umění?
Julie Šišková, Noemi Chrásková, Laura Eklová, Filip Děchtěrenko 135

AI and mental health
Martin Takáč 138

Trade-offs between energetic cost and excitability in compartmental models of juvenile rat CA1 pyramidal cells
Matúš Tomko, Peter Jedlička 143

Genetic Algorithm for Agent Hyperparameter Optimisation in the Test of Algorithmic IQ
Ondřej Vadinský, Michal Dvořák, Matyáš Dlouhý 147

How ShouldWe Teach Robots? A Comparison of Kinesthetic, Joystick, and Gesture-Based Teaching
Petr Vanc, Jan Kristof Behrens, Václav Hlaváč, Karla Štěpánová 156

Could Personhoods Live Symbiotically Within Other Personhoods?
Jiří Wiedermannn 163

Sensory and Emotional Distress as Predictors of Sexual Functioning in Adults with ADHD Traits: A Three-Group Cross-Sectional Study
Anastasiia Yarzhemska 167

Ethical and Psychological Aspects of the Use of Generative AI by Junior Doctors
Karina Zamrazilová 174

Large Behavior Models: a promising approach for safe and reliable robotics?
Branislav Zigo, Igor Farkaš 178

From Remote Control to Vision-Language-Action: Initial Deployment of the Humanoid Robot Unitree G1 with Tactile Hands
Branislav Zigo 187

Author Index 195

Invited speakers

Prof. Jamin Pelkey



Prof. Jamin Pelkey, PhD is Full Professor in the Department of Languages, Literatures & Cultures at Toronto Metropolitan University and a core member of the TMU–York University Joint Graduate Program in Communication & Culture. He currently serves as President of the International Association for Cognitive Semiotics (2022–2026) and previously served as President of the International Cognitive Linguistics Association (2023–2025).

His research focuses on questions of language and meaning and employs mixed methods to investigate the relationships between patterns of embodied experience and the evolution of human consciousness—from extreme ideologies to the creative imagination. His work integrates insights from linguistics, poetics, anthropology, psychology, and philosophy, interpreting symbolic systems across cultures in order to better understand cross-cultural paradoxes and reversals related to belief, knowledge, aesthetic experience, and the meaning of life. Prof. Pelkey is an award-winning teacher and researcher, recipient of multiple prestigious research grants and academic honors, and the author of numerous influential publications in semiotics, cognitive science, and linguistics.

With talk: ***Alteroception and the Divided Brain: Embodied Blending from Imagination to Ideology***

When we experience another person's movement as if we were them, this is known as alteroception. Projections of such experiences onto inanimate spaces provide meaningful networks of criss-crossing relations that owe their dynamic structure to the evolution of upright posture. A series of cross-cultural, trans-historical examples of these patterns at work offer insight into the evolution of language and the origins of human consciousness from imagination to ideology. Recent studies in the neuroscience of brain lateralization suggest that the interplay between these two facets of human nature grows out of radically distinct ways of paying attention to the world. I argue that it is necessary to consider all of these relationships interdependently in order to make sense of human cognition and its consequences.

Dr. Sarah Berger



Dr. Sarah Berger is a Full Professor in the Department of Psychology at the College of Staten Island and a core faculty member at the Graduate Center of the City University of New York (CUNY). She currently serves as the Director of the Child Development Lab at CSI, where she mentors the next generation of developmental scientists. A distinguished scholar with a global research footprint, Dr. Berger is a two-time Fulbright Scholar, most recently, at the Czech Technical University in Prague with Dr. Matej Hoffmann.

Her research focuses on the dynamic interplay between motor action and cognitive processing, employing mixed methods to investigate how infants acquire problem-solving skills in the context of locomotion. Her work integrates insights from developmental psychology, movement science, and artificial intelligence, utilizing interdisciplinary collaborations in computer vision and robotics to model embodied intelligence and automate behavioral coding. She is the author of numerous influential publications in leading journals such as Child Development, Developmental Psychology, and Infancy, contributing significantly to the fields of cognitive science and developmental psychology.

With talk: ***The Role of Sleep in Infant Learning and Problem Solving***

Sleeping is the primary brain activity of early development, yet studies that address how sleep impacts the acquisition of new skills in infancy are scarce. Underscoring the interplay between different development domains, especially at times of emerging new skills, I will present a series of studies from my lab that bridge cognitive development, motor development, and sleep. I will discuss how individual differences in the quality of night sleep, the presence and timing of daytime naps, and gestational age impact infants' motor problem solving. I will also discuss what the temporal relation between the onset of new motor skills and changes in infants' sleep reveals about the function of sleep for learning in infancy.

Mgr. Jindřich Libovický, Ph.D.



Mgr. Jindřich Libovický, Ph.D. is a researcher at the Institute of Formal and Applied Linguistics at MFF UK in Prague. He previously worked at the University of Munich and completed research internships at Google and IBM. At MFF UK, he leads a group focused on multilingual language modeling and cross-lingual fairness. His research interests center on the technical foundations of language representation in language models and on how these models behave across languages – whether their behavior can be predicted from the models' internal states and potentially steered in desirable directions.

With talk: ***How Language Models (Don't Always) Work Across Languages***

Large language models appear at first glance to work well regardless of the language we use to communicate with them. Upon closer evaluation, however, it becomes clear that they perform significantly better for some languages than for others. In this talk, we will explore how multilinguality in language models actually works: How well or poorly do they perform across different languages? To what extent does the amount of available training data for a given language matter? And is it true that language models internally operate in English? We will also show how seemingly unrelated technical decisions made decades ago about character encoding affect model performance – and why these decisions benefit European languages while putting languages of the Global South at a disadvantage.

Mgr. Michal Vavrečka, Ph.D.



Mgr. Michal Vavrečka, Ph.D. is a cognitive scientist specializing in developmental robotics, artificial intelligence, and cognitive modeling. Since the beginning of his research career in 2005, he has been dedicated to the principles of Artificial General Intelligence (AGI). During his psychology studies, he began developing his own cognitive architectures. He continued this work by testing these architectures on humanoid robots trained according to the principles of developmental robotics, using reinforcement learning and intrinsic motivation, and employing advanced robotic simulators that enable the automated generation of long task sequences for lifelong learning. In his research, he integrates findings from psychology and neuroscience with technical applications in robotics. He currently works as a researcher at the Czech Institute of Informatics, Robotics and Cybernetics (CIIRC), Czech Technical University in Prague (CTU), where he focuses on the development of algorithms for humanoid robots and research in Human–Robot Interaction (HRI).

With workshop: *How to Train a Robot*

This workshop provides a hands-on introduction to working with the open-source tool myGym, which enables rapid prototyping of robotics tasks based on reinforcement learning. myGym is a modular simulator designed for training robots in manipulation and navigation tasks. The workshop focuses primarily on hands-on demonstrations: participants will learn how to run a simulation, define a training experiment, and train a robot using reinforcement learning algorithms. Using specific examples, we will demonstrate how to select a robot, environment, and task type, and how to configure the reward function that controls the learning process. myGym allows tasks to be defined parametrically—the user selects the robot, objects, action type, and reward structure, while the training and evaluation themselves are automated. The goal of the workshop is to provide researchers and students with practical guidance on how to quickly begin experimenting with reinforcement learning in robotics, how to design their own experiments, and how to use simulation to develop algorithms that can subsequently be transferred to real robotic systems.

Modeling and Exploring Visual Concepts in Deep Networks

Mhd Walid Al Jallad and Kristína Malinovská

Centre for Cognitie Science, DAI FMPI,
Comenius University in Bratislava, Slovakia
Mlynská dolina, 84248 Bratislava
Email: jallad1@uniba.sk, kristina.malinovska@fmph.uniba.sk

Abstract

Deep neural networks produce powerful visual features, yet they only provide limited capacity for interpretability, compositional reasoning, topographic coherence, and often do not allow for continual learning. Despite their performance, their representations remain far from human cognition and these gaps persist in recent foundation models. Inspired by cortical topographic maps and prototype theory, we propose a differentiable, topology-preserving conceptual layer over features of convolutional and foundation models as a unified interface for visual concept modeling, together with an auxiliary unsupervised loss to guide representations in deep models. Our approach enhances existing models with (1) a navigable concept atlas whose semantic topography mirrors inferotemporal-cortex organization; (2) compositional fuzzy reasoning over concept prototypes, where logical operations respect graded membership and neighborhood smoothness; and (3) continual open-world detection with incremental category learning. The proposal bridges prototype interpretability, concept bottlenecks, neuro-symbolic reasoning, and continual learning under a single biologically grounded computational framework.

1 Introduction

Deep neural networks have become a dominant approach to visual perception, where recent self-supervised foundation models, such as DINOv2/v3, SigLIP, CLIP, are able to produce remarkably strong general-purpose features without requiring labeled data (Oquab et al., 2024; Siméoni et al., 2025; Zhai et al., 2023; Radford et al., 2021). Despite being transferable across distributions and tasks, these features are organized by the training objective rather than by anything resembling a concept. They afford no human-interpretable readout, no compositional reasoning interface, and no principled mechanism for absorbing new categories without forgetting existing ones.

We argue that deep neural networks (DNN) are missing a topographic, prototype-based structure that learns from raw feature spaces in an unsupervised manner. We propose a new approach to tackle the visual concepts within the realm of deep learning inspired by cognitive science methodology. For the sake of our work the visual concept is a labeled category of visual data that the network is driven to classify.

In order to explore the concepts of DNNs and promote emergence of more human-like concepts, inspiration can be taken from biological cognitive systems. First, biological learning combines both bottom-up error-driven and top-down unsupervised processes, as demonstrated by O'Reilly's XCAL model and Leabra framework (O'Reilly et al., 2024).

Second, primate inferior-temporal cortex are shown to contain a coarse navigable map of object space (Bao et al., 2020), suggesting that cortical visual representations are explicitly topographic, a property that DNN feature spaces lack by default. In support of this claim, recent artificial neural networks where able to reproduce this organization only when an explicit topographic prior is added during training (Margalit et al., 2024).

Third, it has been suggested that natural categories in the brain are prototype-like and hierarchically graded (Rosch, 1978), and recent neuroimaging research shows that prototype and exemplar representations coexist across distinct cortical regions during learning (Bowman et al., 2020). This suggests that prototype-based categories behave as fuzzy sets, affording logical composition. Finally, the mammalian brain exhibits efficient lifelong learning due to its reliance on hippocampal episodic memory interleaved with slow neocortical consolidation (McClelland et al., 1995).

Recently, the missing topographic structure in DNNs is slowly starting to get acknowledged by some foundation-models. For instance, the prototype heads of the DINO family have been shown to be redundant and unordered, with many prototypes converging on near-identical directions (Govindarajan et al., 2024). Additionally, DINOv3 introduces an ad-hoc spatial coherence prior to recover the kind of topographic smoothness that classical biologically-motivated representations posses by construction (Siméoni et al., 2025).

Founded in our previous claims, we propose that a topology-preserving prototype layer, i.e., a differentiable self-organizing map (SOM, Kohonen, 1990) defined over a network's feature space, can serve as a uni-

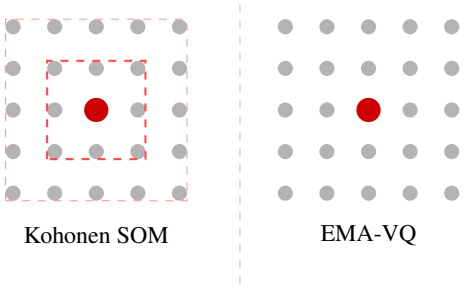


Fig. 1: The Kohonen self-organizing map update with Gaussian neighborhood (left) reduces to the EMA-VQ codebook update under zero neighborhood width (right). The red node marks the best-matching unit; dashed rectangles indicate the neighborhood support. Inspired by (Irie et al., 2024).

fying interface for visual concept modeling. We demonstrate proof of concept by drawing on prior work on an unsupervised auxiliary loss (MT-SOM, Samporová, 2024), and we sketch three downstream affordances: (1) a global, navigable concept atlas with human-validated interpretability; (2) compositional fuzzy-logical reasoning over part-prototypes for top-down concept blending; and (3) continual open-world detection paired with associative memory.

2 Related Work

2.1 Self-organizing Maps as a Topology-preserving Prototype Layer

The classical primitive for topology-preserving prototype learning is the self-organizing map (Kohonen, 1990). A SOM is a two-dimensional grid of prototype vectors trained by repeated competitive selection of a best-matching unit and a neighborhood-weighted update which preserves the topological structure by ensuring prototypes that are close on the grid become close in the feature space. SOMs also exhibit density awareness, in that codebook (i.e. prototype vectors) density tracks the input density.

Recent work has shown that the exponential moving average vector-quantization (EMA-VQ) codebook update used in modern image tokenizers is a special case of the SOM update when the neighborhood approaches zero (Irie et al., 2024). Conversely, the SOM is a strict generalization of the discretization primitive that foundation tokenizers already deploy, since it adds a smooth continuous topological structure without sacrificing the discretization behavior (see Fig. 1)

The integration of SOMs into modern deep learning architectures is a small but a growing research direction. SOM-VAE (Fortuin et al., 2019) and DPSOM (Manduchi et al., 2020) embed SOMs in autoencoders for interpretable time-series representations; SOM-CPC adds SOMs to contrastive learning for high-rate physiological signals (Huijben et al., 2023); and ViT-SOM incorporates a SOM bottleneck into a vision transformer (Luo and Yuan, 2026).

2.2 MT-SOM Model

A concrete instantiation of the SOM mechanism at small scale appears in a prior work on a MT-SOM model with unsupervised auxiliary loss (Samporová, S. and Malinovská, K., 2023; Samporová, 2024). The Mean Teacher model (MT, Tarvainen and Valpola, 2017) maintains a student network and a teacher network whose parameters are an exponential moving average over the student's network. It adds to the supervised loss a *consistency cost* between student and teacher predictions on unlabeled inputs. In MT-SOM (Fig. 2), this consistency cost is reformulated as a distance over a self-organizing map laid over the last convolutional layer of the network. Thus, minimizing the overall loss requires both that student and teacher features map to nearby prototypes on the SOM grid, and that the features themselves remain close to those prototypes.

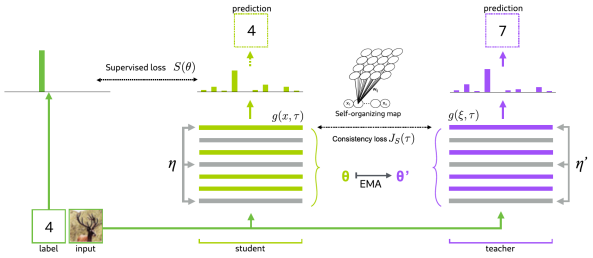


Fig. 2: Illustration of MT-SOM model, figure taken from Samporová (2024)

On CIFAR-10, MT-SOM produces a clearly organized 10×10 prototype grid in which semantically similar classes settle on neighboring grid cells (Fig. 3), and yields modest accuracy gains over the unaugmented MT baseline (Samporová, 2024). The result is small in scale but illustrates the topology preservation and density-aware prototype placement properties of our proposed primitive.

2.3 Four Approaches in Visual Concept Modeling

Four lines of recent work address related aspects of the visual concept modeling problem. The first is regarding prototype-based methods. ProtoPNet (Chen et al., 2019) and its descendants attach a small set of trainable prototypes to a deep neural classifier, and explain predictions as similarity to prototypes. However, recent work shows that spatial alignment of pototype activations to image regions they intend to explain is poor, and that prototype interpretability methods provide no

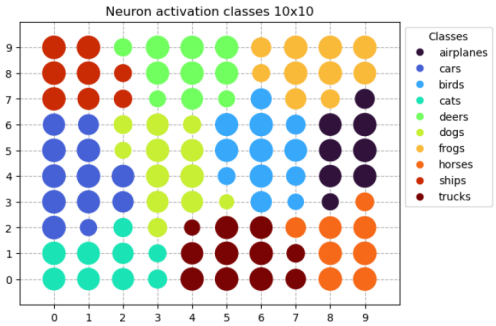


Fig. 3: CIFAR-10 MT-SOM neuron-activation grid, from (Samporová, 2024).

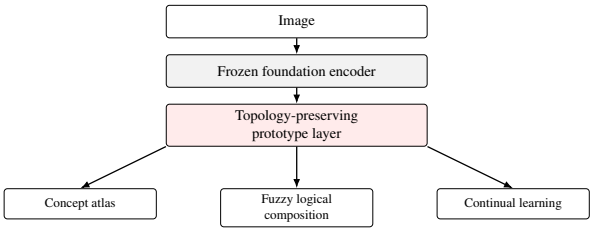


Fig. 4: The proposed pipeline. Images are passed through a frozen foundation encoder; the resulting features are routed through a topology-preserving prototype layer, a differentiable self-organizing map, whose grid serves as a shared substrate for the three downstream affordances.

global geometry over the prototype set (Sacha et al., 2024).

Another instance of a missing global geometry is related to concept-bottleneck models (CBM). Discover-then-Name CBM (Rao et al., 2024) yields a dictionary of named concepts but no relational structure between them, offering per-concept activations rather than a continuous similarity surface. No global geometry of prototype methods are shared, in addition to the risk of leakage when concept activation values are continuous (Mahinpei et al., 2021).

In terms of continual learning, models such as RanPAC (McDonnell et al., 2023) and FeCAM (Goswami et al., 2023) use prototype-style representations for exemplar-free continual learning over frozen foundation features achieving strong results but the prototypes themselves carry no internal organization.

Furthermore, neurosymbolic approaches like PROTO-LTN (Martone et al., 2022) ground logical operations on prototype activations within a logic-tensor-network framework. The logic, however, is imposed without any topological grounding since there is no notion that nearby prototypes should compose more smoothly than distant ones.

3 Our Proposal

We propose a SOM-based mechanism that supports semi-supervised training of DNNs by extending them to include a topology-preserving prototype layer that is trained over the distribution of feature vectors using competitive selection of best-matching units and a neighborhood-weighted update.

This mechanism should be differentiable end-to-end and exhibit three main properties: (1) *topology preservation*—the neighborhood term should ensure that proximity on the grid, representing a conceptual structure, mirrors proximity in feature space; (2) *density awareness*—prototype density tracks the input density in such a way that rare regions of the feature space receive proportionally fewer prototypes; and (3) *discrete sparse activations*, demonstrated by the best-matching unit SOM mechanism, where only a small subset of neurons activate for each input, rather than all neurons being continuously active. Structurally, this is closer to hard concept-bottleneck models than to concept-embedding models that are less transparent due to leakage of information across distributed concept representations (Mahinpei et al., 2021).

To promote the captured structure of the concepts, in line with the previous work on auxiliary unsupervised loss and unsupervised consistency cost (see Sec. 2.2) the models can be further finetuned using the distance on the SOM-based prototype layer. Within the design of the outlined model the map itself shall be very large and sufficiently sparse to avoid collapse of the concepts as well as allow for new concepts to be added post-hoc.

3.1 Downstream Affordances

We sketch three downstream uses of the proposed primitive (see Fig. 4). The first concerns a navigable concept atlas. Existing prototype-interpretability methods do not produce a 2D topology where spatial proximity equals semantic proximity. While each cell in a trained SOM grid carries its prototype vector, which ensures continuous semantic transitions across the grid. This affordance is the most direct translation of the inferior-temporal cortex topographic map (Bao et al., 2020) into a deep learning architecture. To validate such claim, we may adopt the HIVE protocol (Kim et al., 2022), which allows comparing concept-atlas readouts against ProtoPNet-style explanations (Sacha et al., 2024) and Discover-then-Name dictionaries (Rao et al., 2024).

The second affordance concerns compositional fuzzy reasoning. Prototype-based natural categories can be described as fuzzy sets (Rosch, 1978; Bowman et al., 2020), where membership is graded, and compositions of categories, such as AND, OR, NOT, should respect the gradedness. Over the SOM, we predict that nearby part-prototypes should compose more smoothly than distant ones, with neighborhood-weighted t-norms operating on activation patterns. For instance, for a query

"unicorn", the activation pattern should be the intersection of a horse-prototype neighborhood and a horn-prototype neighborhood on the grid. The closet existing comparable model that achieves such compositional reasoning is PROTO-LTN (Martone et al., 2022), which grounds logical operations on prototype activations through logic tensor networks but imposes the logic without any topological structure.

The third affordance is enabling of the continual learning. Recent state-of-the-art continual learning methods that operate over frozen foundation features, use prototype-style representations but offer no mechanism for episodic recall (McDonnell et al., 2023; Goswami et al., 2023). We hypothesize that our topology-preserving prototype layer will offer a natural slow consolidation memory, by initially training a subset of SOM's neurons and freezing the rest in such a way that new categories can either fill in those held-out neurons or settle into existing topographic structure, where rare regions of the grid absorb novel prototypes incrementally.

4 Conclusion and Future Work

In this paper we propose a single primitive, a topology-preserving prototype layer over frozen foundation features, that unifies four lines of recent work in visual concept modeling: prototype interpretability, concept-bottleneck explanation, neurosymbolic reasoning, and continual learning. Each component of our proposal has some inspiration in cognitive neuroscience. Overall, the primitive itself echoes the topographic functional organization of cortical maps (Bao et al., 2020), reflecting the correlation between spatial and semantic representations, and motivating the need of an interpretable and navigable concept atlas that is learned from the feature space of deep neural networks. Fuzzy reasoning draws on prototype theory (Rosch, 1978) and the finding by (Bowman et al., 2020) that prototype and exemplar representations coexist in the brain, motivating composition over graded prototype activations. Finally, continual learning follows complementary learning systems theory (McClelland et al., 1995).

Following this proposal, we will (1) design and benchmark the primitive on deep convolutional neural networks, starting from prior work of MT-SOM (Samporová, 2024) and then expanding the implementation to the attention models, namely the DINO family; (2) compare SOMs navigable concept atlas against other baseline models (e.g., ProtoPNet) utilizing interpretability frameworks; (3) evaluate fuzzy-logic composition over part-prototypes on synthetic compositional benchmarks; and (4) test the SOMs ability on continual learning tasks.

Acknowledgement

This paper was written at the Centre for Cognitive Science at DAI FMPI Comenius University Bratislava, with the support of a project APVV-21-0105. We also thank for support to the Slovak Society for Cognitive SSKV[1].

References

Bao, P., She, L., McGill, M., and Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108.

Bowman, C. R., Iwashita, T., and Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, 9:e59360.

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. (2019). *This looks like that: deep learning for interpretable image recognition*. Curran Associates Inc., Red Hook, NY, USA.

Fortuin, V., Hüser, M., Locatello, F., Strathmann, H., and Rätsch, G. (2019). SOM-VAE: Interpretable discrete representation learning on time series.

Goswami, D., Liu, Y., Twardowski, B., and van de Weijer, J. (2023). FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Govindarajan, H., Sidén, P., Roll, J., and Lindsten, F. (2024). On partial prototype collapse in the DINO family of self-supervised methods. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA.

Huijben, I. A., Nijdam, A. A., Overeem, S., Van Gilst, M. M., and Van Sloun, R. (2023). SOM-CPC: Unsupervised contrastive learning with self-organizing maps for structured representations of high-rate time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14132–14152. PMLR.

Irie, K., Csordás, R., and Schmidhuber, J. (2024). Self-organising neural discrete representation learning à la Kohonen. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 343–362. Springer Nature Switzerland.

Kim, S. S. Y., Meister, N., Ramaswamy, V. V., Fong, R., and Russakovsky, O. (2022). HIVE: Evaluating the human interpretability of visual explanations. In *Computer Vision – ECCV 2022*, pages 280–298. Springer Nature Switzerland.

[1] https://cogsci.fmph.uniba.sk/sskv/

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Luo, A. and Yuan, K. (2026). Simple self-organizing map with vision transformers. *IEEE Signal Processing Letters*, 33:331–335.

Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., and Pan, W. (2021). Promises and pitfalls of black-box concept learning models.

Manduchi, L., Hüser, M., Vogt, J., Rätsch, G., and Fortuin, V. (2020). DPSOM: Deep probabilistic clustering with self-organizing maps.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., and Yamins, D. L. K. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7.

Martone, S., Manigrasso, F., Lamberti, F., and Morra, L. (2022). PROTOtypical logic tensor networks (PROTO-LTN) for zero shot learning. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4427–4433.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev*, 102(3):419–457.

McDonnell, M., Gong, D., Parvaneh, A., Abbasnejad, E., and van den Hengel, A. (2023). RanPAC: Random projections and pre-trained models for continual learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. Featured Certification.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., and Contributors (2024). *Computational Cognitive Neuroscience*. Online Book, 5th Edition, URL: `https://compcogneuro.org`.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Rao, S., Mahajan, S., Böhle, M., and Schiele, B. (2024). Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G., editors, *Computer Vision – ECCV 2024*, pages 444–461, Cham. Springer Nature Switzerland.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*, pages 27–48. Lawrence Elbaum Associates.

Sacha, M., Jura, B., Rymarczyk, D., Struski, Ł., Tabor, J., and Zieliński, B. (2024). Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21563–21573.

Samporová, S. (2024). Auxiliary self-organization-based loss for semi-supervised learning. Master's thesis, Comenius University in Bratislava. Supervisor: K. Malinovská.

Samporová, S. and Malinovská, K. (2023). Využitie samoorganizácie v čiastočne riadenom učení hlbokých neurónových sietí. In Farkaš, I., Ballová Mikušková, E., Takáč, M., Malinovská, K., and Fandl, M., editors, *Kognícia a umelý život 2023*, pages 58–61. Vydavateľstvo UK v Bratislave.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. (2025). DINOv3.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1195–1204, Red Hook, NY, USA. Curran Associates Inc.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Auditory distance perception for sounds around the listener

Gabriela Andrejková and Norbert Kopčo

P. J. Šafárik University in Košice, Faculty of Science
Jesenná 5, 04001 Košice
gabriela.andrejkova@upjs.sk, norbert.kopco@upjs.sk

Abstract

The accuracy of auditory distance perception is very different in the region near the listener ($\leq$1 meter) from the distal region ($>$1 meter). Santarelli et al. (1999) performed an experiment in a reverberant room in which listeners were asked to point to the perceived 3-dimensional (3-D) position of sound sources presented from a random location in the right hemifield within 1 m of the listener's head.

Here, a new analysis focuses on the distance localization performance in these data. It evaluates the mean absolute error (MAE) and standard deviation (SD) in responses as a function of source location varying simultaneously in all three dimensions. Results provide a characterization of 3-D localization performance for nearby sources in a reverberant room (is characterized by high energy density of reflected sound), showing that MAE is significantly different for stimuli (sound sources) directly above the listener compared to stimuli originating in front, behind, or below the listener. SD is highest for stimuli bellow the listener, and it is dependent on distance, polar angle, and lateral angle. Distance judgments are less accurate for medial stimuli than for lateral stimuli at larger distances. The analysis of temporal profile during the whole experiment showed a significant decrease in MAE but not in SD. The decrease in MAE occurred mainly in the second half of the responses, indicating that spontaneous learning has occurred on a slow time on hours.

1 Introduction

The ability to determine the distance of sounds is important for a person's orientation in the environment and for interpretation of these sounds. There are many various acoustical and non-acoustical factors thought to contribute to source distance perception. Zahorik et al. (2005) demonstrated that people use multiple cues to estimate distance, especially the Direct to Reverberant energy Ratio (DRR), systematically underestimate distant sources and these abilities emerge in early childhood.

For sounds near to the listener, Kopčo and Shinn-Cunningham (2011) showed that distance judgments were dominated by a reverberation related cue (commonly interpreted as DRR), which listeners mapped to distance in a direction independent but frequency dependent manner. Brungart and Durlach (1999) evaluated 3-D localization of sound sources near the head in anechoic room (walls absorb sound). They found that in an anechoic space distance localization performance in the proximal field is better than at larger distances, but at the same time it is strongly dependent on azimuth - that is, on whether the source is in a lateral position or in the medial plane.

Santarelli et al. (1999) implemented nearly identical study in a reverberant room and he showed that the presence of reverberation does indeed have a substantial effect on localization ability. The important finding of this work is that different cues underlie the distance perception of nearby sounds in anechoic and reverberant space.

Both studies focused on response analysis separately in each space dimension. Sound localization is typically examined separately in the three spatial dimensions (azimuth, elevation, distance), or for combinations of two dimensions (e.g., azimuth and elevation Best et al. (2011), or azimuth and distance, Ihlefeld and Shinn-Cunningham (1999). Estimating multiple dimensions simultaneously can be more challenging than for fewer dimensions. On the other hand, when location covaries in two dimensions, perception can improve over 1-dimensional baseline (example of visual distance perception Loomis et al. (1998).

Very few studies looked at 3-D localization. Kopčo and Andrejková (2025) examined biases in distance perception for nearby sources varying location in polar and lateral angles in reverberant room. For sources in the horizontal plane the results don't vary dramatically, with an overall underestimation (approximately -10% of presented stimuli distance) that tends to increase for nearby lateral sources (approximately -20 %) and appears to be stronger in front than behind the listeners.

Here, we evaluate distance localization performance of Santarelli et al. (1999) data in a perceptually based coordinate system, in which the space was divided into 9 directional bins based on lateral angles (bin centered at 15°, 45°, 75°), polar angles (bins centered at 0°, 90°, 180°, 270°). Each directional bin is divided into 5 log-distance bins (centered at 20, 28, 41, 58, 83 cm). We use MAE as the main accuracy metric and SD as

the reliability metric. MAE assesses the overall localization error, while SD expresses the variability of the responses.

Together, the two metrics provide a robust assessment of accuracy and consistency in the near field, where acoustic parameters cause nonlinear and significant variations.

2 Material and methodology

2.1 Original data

The experiment was done by Santarelli et al. (1999) in the room (9x5x3-m (length x width x height) located within the Department of Cognitive and Neural Systems at Boston University. The testing apparatus, shown in Fig. 1, was prepared according to description of Brungart and Durlach (1999). Seven listeners participated in the experiment. The listener sat on a wooden chair placed in the center of the room. The chair was equipped with an adjustable headrest to help prevent the listener from making head movements during data collection. The experimenter stood to the right of the listener with the testing apparatus. The transmitter for the electromagnetic tracker was placed to the rear of the listener, slightly to the left, at mid-back level.

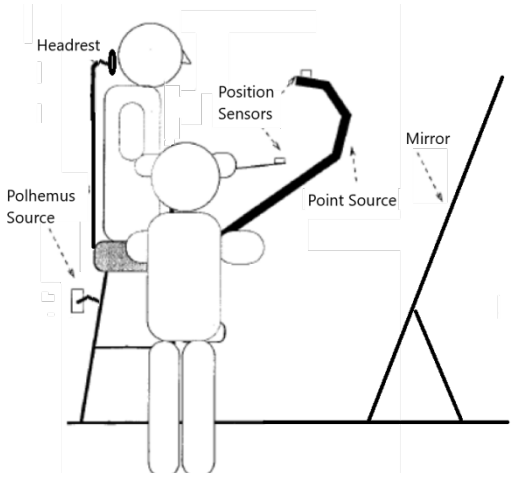


Fig. 1: Modified setup according to (Brungart and Durlach (1999))

2.2 Procedure

At the beginning of each trial, a computer-generated voice instructed the listener to ”listen,” providing the listener with a cue to close his eyes. The computer then randomly chose a spatial location in the listener’s right hemisphere by specifying the appropriate azimuth, elevation, and distance. Next, the experimenter pressed a switch on the source-wand, triggering the presentation of the stimulus. The experimenter promptly returned the source to the starting position and pressed the switch again. The listener opened his eyes and used the response wand to indicate the position at which he perceived the source. With a final press of the switch, this position was recorded, marking the end of the trial.

Stimuli consisted of five pink-noise bursts (150 ms duration, 30 ms inter-stimulus interval, or ISI). During data collection, the trials were partitioned into blocks that consisted of either 25 or 50 trials. Each listener was given a 5–10-minute rest interval between blocks of trials. A typical session lasted roughly 1 to 1.5 hours and included a total of 100 to 150 trials. Each listener was required to complete a total of 1200 trials (the first 200 of which were practice trials), which translates to approximately 12 hours of testing.

2.3 Methodology

We evaluated the results using the interaural polar coordinate system which allowed us to track the dependence of the distance responses simultaneously on the lateral angle (LA) and polar angles (PA), as well as on source distance. The data were binned into 9 directional bins by dividing them into five distances. The directional bins were combinations of lateral angle (3 regular intervals centered at LA = [15°, 45°, 75°]) and polar angle (one bin for LA $>$ 75°, 4 bins centered at PA = [0°, 90°, 180°, 270°] for LA in range of [15°, 45°]). We evaluated Mean Absolute Error (MAE) and Standard Deviations (SD) using a log-log scale (log10(response distance) – log10(stimulus distance)). Log–log scales are used in hearing research because both sound and human perception operate on ratios and power laws, not linear differences. Repeated measures ANOVA were used to analyze the data with Box-Geissler-Greenhouse epsilon used to correct for potential violations of the sphericity assumption.

3 Results

3.1 MAE and SD

We averaged data within each directional and distance bin for each listener and then collapsed across listeners. In the upper panels of Fig. 2, the MAE is shown as a function of log-distance, the polar angle is distinguished by color and a colored icon. The largest deviations are observable in stimuli in bins with a centered polar angle of 270° (magenta), the stimuli below listeners. For LA=15°, near the medial plane, MAE increases with distance for stimuli above the listener. In other directions (front, back, and below) MAE is affected by lateral angle only. Listeners make larger errors in smaller lateral angles, and with increasing lateral angle these errors don’t increase.

A three-way ANOVA with factors polar angle PA = [0°, 90°, 180°, 270°], lateral angle LA = [15°, 45°]

and distance D = [20, 28, 41, 58, 83] cm showed a significant main effect LA, $F(1,6) = 35.32$, $p = 0.001$, indicating that MAE for LA = 15° has higher values than for LA = 45°. At the same time, a significant interaction PA × D was shown, $F(12,72) = 3.49$, $p = 0.0004$. Subsequent simple effects analyses revealed that distance D is mostly increasing with distance at PA = 90° (red line).

In evaluation SD, 3-way ANOVA showed main effects PA, $F(3,18) = 12.47$, $p<0.001$, LA, $F(1, 6) = 17.36$, $p = 0.003$, and D, $F(4,24) = 5.52$, $p = 0.003$. SD is affected by distance, polar and lateral angle independently of each other.

After averaging across polar angles, 2-way ANOVA with factors LA = [15°, 45°, 75°] and D = [20, 28, 41, 58, 83] cm showed main effect LA, $F(2,12) = 64.16$, $p<0.001$.

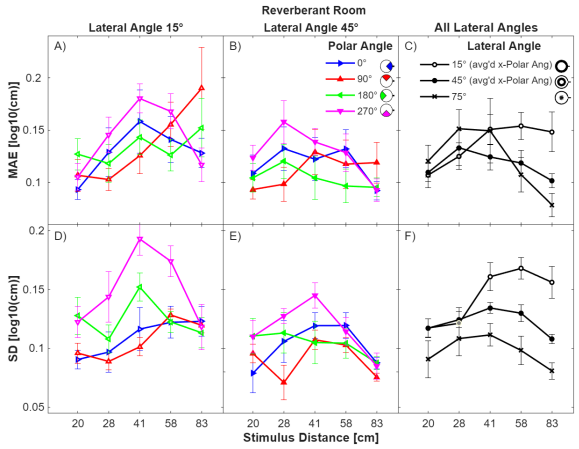


Fig. 2: Across-listener mean of Absolute Error (MAE) and Standard Deviations (SD) in five distance and nine directional bins. The upper panels A) - B) show MAE in two lateral angels and four polar angles (recognized by colors and icons). Panel C) shows the MAE for all three lateral angles after averaging over polar angles. In the lower panels, SDs are drawn in a similar way.

Evaluation MAE and SD in panels C) and F) shows the largest differences for lateral angle LA = 15°. This means the presence of outliers and smaller stability of responses. Listeners' responses to stimuli in this area are the most imprecise and variable.

3.2 Temporal profile

Each listener performed approximately 1200 trials in the same reverberant room. The first 200 trials are considered training stimuli, and they weren't used in an evaluation of temporal profile. In rooms, reverberation provides distance information and candidate cue is Direct-to-Reverberant energy ratio, Bronkhorst and Houtgast (1999). Shinn-Cunningham (2000) showed that in rooms, there exists a learning effect: distance performance is improved with experience.

We analyzed learning process followed by MAE and SD through all evaluated trials. All data of each listener was divided into 5 temporal blocks per 200 trials in chronological order. The sequence of these blocks follows time sequence of trials. For each listener and each trial, the MAE and SD was calculated and averaged across stimuli in each temporal block. The time course of the averaged values also represents the course of adaptation to the stimuli.

In Fig. 3, panel A), the courses of all listeners are shown (colored dotted lines), as well as their average across listeners (black solid line). We assumed some trend in decreasing MAE, a linear approximation of MAE is depicted in panel B), for all listeners and mean across listeners (black solid line). SDs are shown in panels D) and E). One-way ANOVA with factor the first and the last values of approximating line V (V1, V2) showed main effect V $F(1,6) = 6.05$, $p = 0.049$. This means that the MAE is decreasing and the decrease in values is significant.

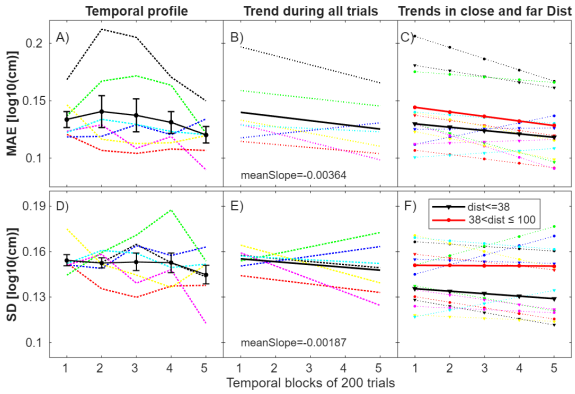


Fig. 3: Temporal Profile (TP) of MAE and SD. The first column of panels A) and D) shows profiles of MAE and SD of all listeners (colored dotted lines) and the mean across the listeners (black solid line) in 5 temporal blocks of 200 averaged trials. The panels B) and D) show linear regressions of temporal profiles. In the panels C) and F), trends of listeners (dotted lines) and averaged across listeners in two log-distances are shown.

Further analysis of a temporal profile in each of two log-distances intervals, division in 38 cm, is shown in panel C). Averaged values across stimuli and across listeners in each log-distance interval are drawn in five temporal blocks. It is shown that MAE doesn't increase in any log-distance interval, in most of them MAE decreases. Two-way ANOVA with factors distance D (≤ 38, > 38) and begin and end values of approximating lines V (V1, V2) showed main effect V, $F(1,6) = 6.11$, $p = 0.048$. It means that the MAE is decreasing and the decreasing in values is significant at V, but there is no significant effect at D. Taken together, these results indicate that listeners continue to improve on the task in both distances even after 200 practice trials.

4 Discussion

The current study examined MAE and SD in distance perception for nearby sources varying location in polar and lateral angles in the reverberant room. For frontal sources close to the horizontal plane (polar angle from -45° to +45°) the results don't vary dramatically (blue lines), the highest value of MAE is in bin centered by 41 cm for frontal sources (panel A). For sources above the listener's head (from 45° to 135°), we observe an increasing MAE with distance for frontal sources. The largest MAE values are for sources below the listener's head at almost all distances. In the regions of lateral angles (0° to 30°), (30° to 60°) and (60° to 90°) we observe an increasing in MAE at small distances, at larger distances the MAE decreases or almost does not change, panel C). The SD is largest for sources below the listener's head and larger for frontal sources, panel D). At larger distances, we observe that SD is almost the same for all polar angles, panels D) and E). Panel F) shows the largest variability for frontal sources.

Although Brungart and Durlach (1999) investigated the localization of very close sources in anechoic space, his results provide a generally valid explanation of the psychophysical and acoustic limitations that affect the accuracy of distance estimation. Brungart shows that for sources up to about 1 m, there are sharp nonlinear changes in binaural and spectral cues, especially the interaural level difference (ILD), while the interaural time difference (ITD) remains relatively constant. These rapid changes in acoustic cues cause impaired distance estimation accuracy and increased response variability already in anechoic environments.

In a reverberant room, these known limitations are compounded by the contribution of DRR, which is a significant distance cue for nearby sources. Santarelli et al. (1999) and later Kopčo and Andrejková (2025) show that reverberation can improve or worsen the availability of distance information depending on the direction of the source, with systematic upward and downward biases being typical of the combination of binaural and reverberant cues.

The resulting MAE and SD values in our data thus represent a natural consequence of these known mechanisms. MAE reflects a combination of the nonlinear changes in ILD, and spectral properties described by Brungart, supplemented by the distribution of reverberation in the room. SD results from the instability of acoustic cues with small changes in source distance and orientation, which are even more pronounced in a reverberant environment. Overall, MAE and SD are therefore consistent with Brungart's theory of local acoustic limitations for very close sounds, as well as with the results of Santarelli et al. (1999) in reverberant conditions.

Our results confirm previous findings Shinn-Cunningham (2000) that room learning occurs in reverberant environments, leading to reduced localization errors. The improvement in MAE and SD during the experiment is therefore an expected and theoretically well-supported consequence of adaptation to the acoustic environment.

Acknowledgement

Work supported by EU HORIZON-MSCA-2022-SE-01 101129903, APVV-23-0054, SK-AT-23-0002.

References

Best, V., Carlile, S., Kopčo, N. a van Schaik, A. (2011). Localization in speech mixtures by listeners with hearing loss. *The Journal of the Acoustical Society of America*, 129(5):210–215.

Bronkhorst, A. W. a Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, 39(11):517–520.

Brungart, D. S. a Durlach, N. I. (1999). Auditory localization of nearby sources. ii. localization of a broadband source in the near field. *The Journal of the Acoustical Society of America*, 106(4):1956–1968.

Ihlefeld, A. a Shinn-Cunningham, B. G. (1999). Effect of source spectrum on sound localization in an everyday reverberant room. *The Journal of the Acoustical Society of America*, 130(1):324–333.

Kopčo, N. a Andrejková, G. (2025). Effect of stimulus spectrum on distance perception for nearby sources. *Proc. Forum Acusticum 2025*, str. 3393–3397.

Kopčo, N. a Shinn-Cunningham, B. G. (2011). Effect of stimulus spectrum on distance perception for nearby sources. *The Journal of the Acoustical Society of America*, 130(3):1530–1541.

Loomis, J. M., Klatzky, R. L., Philbeck, J. W. a Golledge, R. G. (1998). Assessing auditory distance perception using perceptually directed action. *Perception & Psychophysics*, 60(6):966–980.

Santarelli, S., Kopčo, N., Shinn-Cunningham, B. G. a Brungart, D. S. (1999). Near-field localization in echoic rooms. *The Journal of the Acoustical Society of America*, 105(2):1024.

Shinn-Cunningham, B. G. (2000). Learning reverberation: Implications for spatial auditory displays. *Proc. International Conference on Auditory Displays*, str. 126–134.

Zahorik, P., Brungart, D. S. a Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United with Acustica,*, str. 409–420.

Semantic Robustness of Unsupervised Approaches for Interpretability of Vision Language Models

Tamara Bíla, Adam Červenka, Igor Farkaš

Centre for Cognitive Science, Department of Applied Informatics
Comenius University Bratislava, Slovakia
Mlynská dolina, 84248 Bratislava
tamara.bila@fmph.uniba.sk

Abstract

The problem of interpreting deep neural networks, especially large transformer-based models, is widely known and has motivated the development of diverse interpretability approaches. A dominant branch of interpretability research focuses on concept discovery in latent spaces. Sparse Autoencoders (SAEs) and clustering methods are two prominent unsupervised concept analysis approaches. However, little is known about the consistency of the structures recovered. In this work, we propose a framework for evaluating the semantic robustness of these methods by applying SAEs and clustering techniques in parallel to the same latent spaces of the Vision-Language Model (VLM) LLaVA. Our proposed theoretical framwork combines layer-wise overlap measures, MonoSemanticity scores, and visual inspection of representative samples and importance heatmaps. The goal is to investigate whether the discovered concepts reflect the intrinsic semantic structure of the target model or arise as artifacts of the chosen extraction method.

1 Introduction

Modern foundation models are able to process highly rich latent representations, but their structure and the layer-wise progression of activations are hardly interpretable. A promising and popular scheme for understanding latent structure is the so-called concept analysis, which aims to map internal representations to human-understandable semantics. Within the spectrum of available concept extraction methods, clustering and Sparse Autoencoders represent two prominent unsupervised approaches. However, the associated algorithms are fundamentally different.

Clustering groups latent activations according to their similarity, implicitly assuming that semantic structure aligns with geometric proximity. In contrast, SAEs learn sparse, overcomplete feature dictionaries, where individual neurons or directions are interpreted as disentangled concepts. While both approaches are used widely, there is limited understanding of whether they recover consistent semantic structures or whether their outputs are method-dependent artifacts.

In this work, we propose a theoretical framework for investigating the extent to which concepts recovered by different unsupervised interpretability techniques are stable and method-invariant. Specifically, we propose to apply clustering methods and Sparse Autoencoders (SAEs) in parallel to the same latent representations of a foundation model LLaVA Liu et al. (2024). The framework combines a layer-wise overlap measure, MonoSemanticity scores, and visualization of the representative samples to evaluate whether both approaches recover consistent semantic structure. Note that the MonoSemanticity score for SAEs was introduced by Pach et al. (2025). Here, we extend this idea by proposing an analogous metric for clustering-based concepts (Section 3).

The present work focuses on the formulation of the proposed framework and its associated evaluation methodology, while its implementation and large-scale empirical validation constitute the primary direction of future work.

The proposed framework is motivated by the following hypothesis: if concepts extracted by clustering and SAEs converge, this would suggest that the underlying semantic structure is intrinsic to the model. Conversely, systematic disagreement between the two approaches would indicate that at least part of the recovered structure may arise from biases introduced by the chosen extraction method.

2 Related Work

Concept-based interpretability aims to associate latent representations of neural networks with human-understandable semantic concepts. Existing approaches can be broadly/ divided into supervised and unsupervised methods.

Supervised methods typically rely on predefined concept annotations. Concept Bottleneck Models (CBMs) Koh et al. (2020); Rao et al. (2024) explicitly predict concepts before producing final outputs, while Concept Whitening Chen et al. (2020) aligns latent directions with semantic concepts through fine-tuning a Concept Whitening module. Another influential post

hoc approach is TCAV Kim et al. (2018); Schrouff et al. (2022), which estimates concept activation directions in latent space and measures their importance with respect to model predictions. Since TCAV assumes approximate linear separability of concepts, Crabbé and van der Schaar (2022) proposed nonlinear support vector classifiers to better capture complex latent geometries. Although effective, supervised approaches generally require a datasets annotated with predefined concepts, restraining scalability and generalization.

These limitations motivated increasing interest in unsupervised concept discovery methods. Sparse Autoencoders (SAEs) Huben et al. (2024) learn sparse overcomplete feature dictionaries and are commonly interpreted through the superposition hypothesis Elhage et al. (2022), where individual neurons correspond to disentangled semantic directions. However, SAE training often requires very large datasets, and interpretation of the learned features frequently depends on external language or vision-language models Lieberum et al. (2024); Zhang et al. (2025).

A complementary line of work employs clustering techniques to uncover semantic organization in latent spaces. Prior work demonstrated the effectiveness of unsupervised clustering approaches for discovering meaningful latent structures in both vision and language models Kolouri et al. (2017); Ghorbani et al. (2019); Dalvi et al. (2021). In particular, O'Mahony et al. (2023) showed that hierarchical clustering can disentangle multiple semantic meanings hidden within polysemantic neuron activations. Fel et al. (2023) further proposed a unifying framework interpreting concept extraction as a form of dictionary learning, while Hawasly et al. (2024) analyzed quality–efficiency trade-offs of clustering-based concept discovery methods.

Despite the growing popularity of both SAEs and clustering approaches, relatively little work investigates whether the semantic structures recovered by these methods are consistent across extraction paradigms. Our work addresses this gap by directly comparing clustering-based and SAE-derived concepts within the same latent spaces and by evaluating their robustness under varying data conditions.

3 Methodology

In order to test the robustness of the unsupervised concept extraction approaches, we state a proposition. Particularly, if concepts are encoded in the latent space – manifold – of the target model, they should be discoverable irrespective of the method employed, assuming the used method is faithful to the latent representations. To this end, we implement and test two interpretability approaches, namely clustering-based and SAE concept learning. Specifically, we chose and apply both methods to subset of deep layers in a pretrained Vision-Language Model, such as LLaVA.

Then, we compare the semantic similarity between clusters and SAE features measured by their pointwise overlap, given as the weighted cosine similarity. For the normalized feature activations of the k-th SAE neuron, $\tilde{a}_n^k \in [0,1]$ and the smoothed cluster membership $\tilde{c}_n^j \in [0,1]$ given by the normalized angular distance to j-th centroid the overlap becomes

$$O_{jk} := \frac{\sum_{n=1}^{N} \tilde{c}_n^j \tilde{a}_n^k}{\sqrt{\sum_n (\tilde{c}_n^j)^2 \sum_n (\tilde{a}_n^k)^2}} \tag{1}$$

Note that both SAE feature activations and the cluster memberhip of latent activations refer to the same layer ℓ from which they are extracted. In addition, all activations come from a common dataset $\mathcal{J} = \{\mathbf{x}_n \in \mathbb{X}\}_{n=1}^{N}$, which should span a wide variety of images to include enough concept of diverse granularity. Alternatively, a probability based measure can be defined using the Jensen-Shannon divergence (JSD), which is a symmetrized version of the Kullback-Leibler divergence for two distributions by transforming the normalized activations into probabilities $p_k(n) = \tilde{a}_n^k / \sum_m \tilde{a}_n^k$.

A high overlap between a pair of cluster and SAE neuron indicates a substantial semantic closeness.

The validity of semantic coherence of each SAE feature and cluster can also be tested individually using the MonoSemanticity score introduced by Pach et al. (2025). For the SAE features the score of the k-th neuron becomes

$$\mathrm{MS}_{\text{sae}}^k = \frac{\sum_{1 \le n < m \le N} r_{nm}^k s_{nm}}{\sum_{1 \le n < m \le N} r_{nm}^k} \tag{2}$$

where the relevance matrix $r_{nm}^k = \tilde{a}_n^k \tilde{a}_m^k$ quantifies the joint neuron activation of each image pair. s_{nm} is the pairwise similarity between images computed as the cosine distance of their embeddings from an auxiliary vision encoder E, i.e.

$$s_{nm} = \frac{E(\mathbf{x}_n) \cdot E(\mathbf{x}_m)}{|E(\mathbf{x}_n)||E(\mathbf{x}_m)|}. \tag{3}$$

Note that in future experiments CLIP network (Radford et al., 2021) will be used as the vision encoder E. We formulate an analogous MonoSemanticity metric adapted to clusters, which replaces neuron activations $\tilde{a}_n^k$ by the normalized distance of each activation $\mathbf{a}_n$ in the embedding space from the cluster centroid $\boldsymbol{\mu}_j$:

$$\tilde{d}_n^j = \frac{1 + \cos(\theta)}{2} \quad \text{with} \quad \cos(\theta) = \frac{\mathbf{a}_n \cdot \boldsymbol{\mu}_j}{|\mathbf{a}_n||\boldsymbol{\mu}_j|}. \tag{4}$$

Then the relevance matrix is computed analogously as for the SAE metric, that is, $\rho_{nm}^j = \tilde{d}_n^j \tilde{d}_m^j$. Hence, applying the same logic to construct the MonoSemanticity score for clusters yields

$$\mathrm{MS}_{\text{cl}}^j = \frac{\sum_{1 \le n < m \le N} \rho_{nm}^j s_{nm}}{\sum_{1 \le n < m \le N} \rho_{nm}^j}. \tag{5}$$

Computing the aforementioned metrics allows a quantitative comparison of qualitatively distinct concept extraction methods. On the one hand, the SAEs are trained in a self-supervised fashion, disentangling the latent signal into an overcomplete dictionary of features using an encoder-decoder like architecture. On the other side, the clustering-based approaches decompose the latent activations into groups (clusters) according to their pairwise (agglomerative clustering) or point to centroid (K-means clustering) distances. Therefore, finding matching concepts among the SAE features and the clusters strengthens the validity of the individual results and affirms the robustness of both present concepts extraction approaches.

4 Visual Inspection Techniques

Once the overlap between SAE features and clusters is established, the feature-cluster pairs with overlap higher than a set threshold $O_{jk} \geq \tau$ will be selected for introspection and concept description. The overlap threshold τ may be selected statistically by estimating a null distribution through random permutation of cluster memberships or SAE activations and choosing the corresponding upper confidence percentile (e.g., the 95th percentile), thereby retaining only statistically significant concept correspondences. Presumably, clusters and features whose activations overlap significantly share common input samples and will thus constitute the same (or similar) semantic meaning – concept. Particularly, samples whose activations fire for both a feature and a cluster simultaneously can be considered as the core representatives for the sought-for concept.

Conversely, examining the semantic content of clusters and SAE features independently may reveal discrepancies between the two approaches, thereby exposing inherent differences in the corresponding extraction methods. These differences constitute an important object of study in their own right.

To interpret these features, it is necessary to determine which semantic concepts they embody and how these concepts can be mapped to human language. At present, there exist two primary approaches for inspecting and describing their semantic content in natural language. The first, most intuitive yet labor-intensive approach relies on human visual inspection of the corresponding input images. To illustrate this method, we provide an example below. Another possibility is to employ an external Vision-Language Model (VLM) capable of generating textual descriptions for sets of images associated with a given cluster or SAE feature. In this work, we do not further explore the latter approach, although it may serve as an interesting direction for future analysis.

Here, we briefly present the process of visually inspecting the input images associated with the highest activations a_n^k of the k-th neuron in the SAE. Similarly, the semantic content of a cluster can be analyzed by displaying the input samples whose activation vectors $\mathbf{a}_n$ lie closest to the centroid $\boldsymbol{\mu}_j$ of a given cluster. Furthermore, importance heat maps are constructed for a subset of representative samples in order to highlight the input patches that contribute most strongly to the activation of the corresponding neuron.

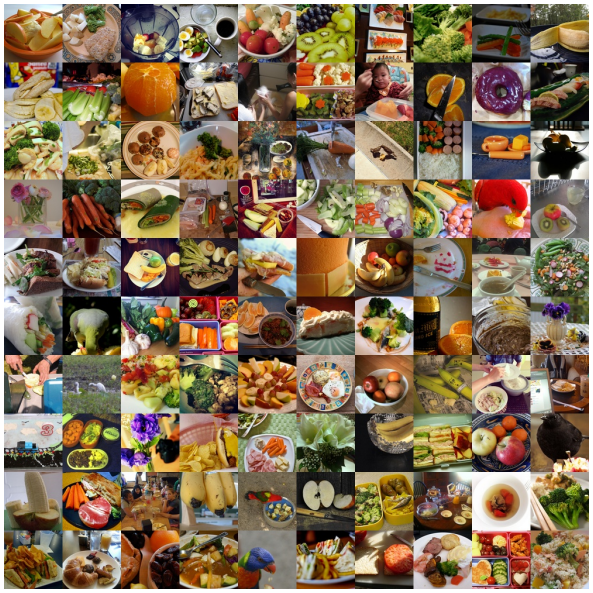


Fig. 1: A collage plot of 100 input images, which yield the highest activation of the 41045th neuron in the SAE layer. The apparent common concept among these images is *food* dominated by fruits and vegetables.

A composite visualization of the 100 most representative samples associated with the 41 045-th SAE feature, trained on the 18-th layer of LLaVA 1.5 Liu et al. (2024), is shown in Figure 1.

Another useful approach, mentioned above, is the construction of importance maps or heatmaps. Figure 2 presents four heatmaps highlighting the input patches that contribute the most strongly to the activation of the same 41 045-th neuron. Interestingly, the boundaries of the depicted objects, predominantly fruits and vegetables, consistently exhibit the highest contributions to the feature activation.

5 Conclusion and Future Work

To conclude, we have proposed a framework together with quantitative measures for evaluating the robustness of extracted concepts with respect to the choice of unsupervised extraction method. The proposed methodology consists of three main stages. First, Sparse Autoencoders are to be trained on selected layers of a foundation model, while activations from the same layers are analyzed in parallel using clustering techniques. Next, the pairwise overlap between SAE features and clusters



Fig. 2: Importance heatmap revealing image patches with highest contribution to the SAE feature activation. The four presented images belong to the set of samples with overall highest activation of the 41045th neuron.

is evaluated layer-wise, together with MonoSemanticity scores computed for the top-N activating samples and the top-N samples closest to the corresponding cluster centroids. Finally, representative samples associated with both SAE features and clusters are visualized using collage plots, complemented by importance heat maps highlighting the input regions contributing most strongly to feature activation.

The primary direction of future work is the implementation and large-scale evaluation of the proposed framework. The experimental analysis will be conducted on a large and semantically diverse image dataset (e.g., ImageNet, COCO, or Visual Genome) ensuring that the extracted concepts span a broad spectrum of visual semantics and abstraction levels. The expected contribution lies in providing empirical evidence regarding the semantic robustness of unsupervised concept discovery methods. In particular, the proposed analysis aims to clarify the relationship between geometric, clustering-based representations and sparse SAE-derived representations. Quantitative overlap and monosemanticity measures may provide practical insight into the reliability of these methods and help determine in which settings one approach may be preferable over the other.

More broadly, improving our understanding of the stability and consistency of unsupervised interpretability techniques may contribute to more reliable and trustworthy deployment of foundation models.

Acknowledgement

This work was supported by the TRAIL project (TRAnsparent, InterpretabLe Robots), funded by Horizon Europe (HORIZON) as a Marie Skłodowska-Curie Action Doctoral Network (MSCA-DN) under the grant agreement No. 101072488.

References

Chen, Z., Bei, Y. and Rudin, C. (2020). Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2:1–11.

Crabbé, J. and van der Schaar, M. (2022). Concept activation regions: A generalized framework for concept-based explanations. In *Neural Information Processing Systems*.

Dalvi, F., Khan, A. R., Alam, F., Durrani, N., Xu, J. and Sajjad, H. (2021). Discovering Latent Concepts Learned in BERT. In *International Conference on Learning Representations*, Virtual Event, Austria. Poster at International Conference on Learning Representations (ICLR) 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M. and Olah, C. (2022). Toy models of superposition.

Fel, T., Boutin, V., Moayeri, M., Cadène, R., Bethune, L., Andéol, L., Chalvidal, M. and Serre, T. (2023). A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ghorbani, A., Wexler, J., Zou, J. Y. and Kim, B. (2019). Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

Hawasly, M., Dalvi, F. and Durrani, N. (2024). Scaling up Discovery of Latent Concepts in Deep NLP Models. Graham, Y. and Purver, M. (ed.), In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 793–806, St. Julian's, Malta. Association for Computational Linguistics.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A. and Sharkey, L. (2024). Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *The Twelfth International Conference on Learning Representations*, pp. 7827–7845, Vienna, Austria. ICLR.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. and Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B. and Liang, P. (2020). Concept bottleneck models. III, H. D. and Singh, A. (ed.), In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 from *Proceedings of Machine Learning Research*, pp. 5338–5348, Virtual, Austria. PMLR.

Kolouri, S., Martin, C. E. and Hoffmann, H. (2017). Explaining Distributed Neural Activations via Unsupervised Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1670–1678, Honolulu, HI, USA. IEEE.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R. and Nanda, N. (2024). Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. Belinkov, Y., Kim, N., Jumelet, J., Mohebbi, H., Mueller, A. and Chen, H. (ed.), In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP of Association for Computational Linguistics,*, pp. 278–300, Miami, Florida, US. Association for Computational Linguistics.

Liu, H., Li, C., Li, Y. and Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306.

O'Mahony, L., Andrearczyk, V., Müller, H. and Graziani, M. (2023). Disentangling neuron representations with concept vectors. In *Conference on Computer Vision and Pattern Recognition*, pp. 3770–3775. IEEE.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S. and Akata, Z. (2025). Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR.

Rao, S., Mahajan, S., Böhle, M. and Schiele, B. (2024). Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G. (ed.), In *Computer Vision – ECCV 2024*, pp. 444–461, Cham. Springer Nature Switzerland.

Schrouff, J., Baur, S., Hou, S., Mincu, D., Loreaux, E., Blanes, R., Wexler, J., Karthikesalingam, A. and Kim, B. (2022). Best of both worlds: local and global explanations with human-understandable concepts. arXiv:2106.08641.

Zhang, K., Shen, Y., Li, B. and Liu, Z. (2025). Large multi-modal models can interpret features in large multi-modal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3650–3661, Honolulu, Hawaii, US. IEEE.

Zkreslení pozornosti při percepci jídla: Eye-trackingová studie obézních dětí

Jiří Čeněk [a*], Stanislav Mokrý [b], Lenka Hromková [a], Beatrice Elena Chromková Manea [a], Leona Mužíková [c], Veronika Suchodolová [c], Anna Svozilová [a], Dana Hübelová [a]

[a] Fakulta regionálního rozvoje a mezinárodních studií, Mendelova univerzita v Brně, Zemědělská 1665/1, 613 00 Brno, Česká republika
[b] Provozně ekonomická fakulta, Mendelova univerzita v Brně, Zemědělská 1665/1, 613 00 Brno, Česká republika
[c] Lékařská fakulta, Masarykova univerzita, Žerotínovo nám. 617/9, 601 77 Brno, Česká republika

[*]korespondenční autor: jiri.cenek@mendelu.cz

Abstrakt

Výzkumy naznačují, že u obézních lidí dochází ke zkreslení vizuální pozornosti směrem ke stimulům spojených s jídlem, oproti neobézním lidem již na úrovni percepce věnují pozorování jídla více času. V této eye-trackingové studii porovnáváme míru pozornosti věnovanou různým typům vizuálních podnětů (vysoce a nízce průmyslově zpracované jídlo, ne-jídlo) na vzorku dětí z léčebny pro obézní děti a kontrastní skupině neobézních dětí. Eye-trackingová úloha sestávala ze dvou částí. V první části respondenti volně pozorovali kombinace dvou typů vizuálních podnětů. Ve druhé části byla část podnětů obsahující kombinaci vysoce a nízce průmyslově zpracované jídlo prezentována znovu a respondenti si vybírali jídlo, na které mají více chuť. Mimo experimentální část byl administrován dotazník DEBQ-C (Dutch Eating Behavior Questionnaire). Výsledky naznačují, že u obézních dětí z léčebny ve srovnání s kontrolní skupinou dochází k opačnému efektu, než je popisován v literatuře, tedy že vysoce zpracovanému jídlu věnují méně pozornosti. Míra pozornosti věnovaná jídlu souvisí se subškálami DEBQ-C.

1 Úvod

Obezita v dětství je komplexní fenomén, jehož výskyt v posledních desetiletích celosvětově roste (Wang & Lobstein, 2006). Vedle genetických a environmentálních faktorů hrají důležitou roli i kognitivně-behaviorální mechanismy, jako jsou stravovací návyky a pozornostní procesy.

Podle teorie předchozích výzkumů na zkreslení pozornosti při percepci jídla, může vystavení podnětům spojeným s jídlem vést k větší citlivosti na tyto podněty, a tedy i k větší pravděpodobnosti nadměrného příjmu kalorií. Výzkumy u dospělých i adolescentů často naznačují, že jedinci s vyšší tělesnou hmotností věnují větší pozornost vysoce kalorickému jídlu. Nicméně tato zjištění nejsou u dětí konzistentní (Graham et al., 2011; Werthmann et al., 2011).

Naším cílem bylo ověřit, zda existuje rozdíl v délce fixací a preferencích mezi dětmi s obezitou a jejich vrstevníky s normální hmotností. Zároveň jsme chtěli zjistit, zda existuje vztah mezi stravovacím stylem (měřeným pomocí DEBQ-C) a pozorovanými pozornostními vzorci.

2 Metody

2.1 Výzkumný soubor

Do studie bylo zahrnuto 77 dětí ve věku 11 až 17 let. Skupina dětí s obezitou (n = 43) byla rekrutována v rámci lázeňského pobytu zaměřeného na léčbu nadváhy. Kontrolní skupinu (n = 34) tvořily děti bez obezity, které se rekreačně zotavovaly v jiném zařízení. Obě skupiny byly srovnatelné z hlediska věku a pohlaví. Přesná antropometrická kritéria, například BMI percentil nebo BMI z-skóre, nebyla v této fázi analýzy zahrnuta

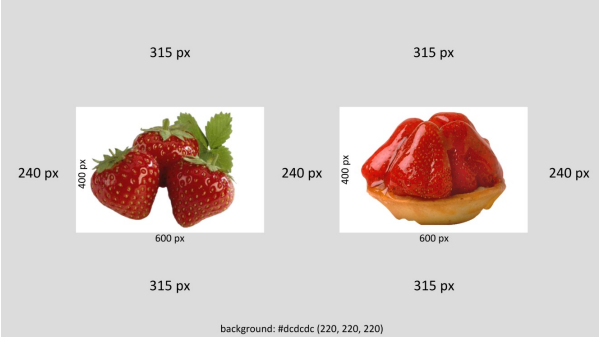
2.2 Materiály

K měření stravovacích návyků byl použit dotazník DEBQ-C (Dutch Eating Behavior Questionnaire for Children), který zahrnuje tři dimenze: restriktivní, emoční a externí stravování (van Strien, et al., 2008). Původní binární škály odpovědí v dotazníku byly upraveny na pětibodové frekvenční Likertovy škály: „nikdy“ (0), „zřídka“, „někdy“, „často“ a „velmi často“ (4).

Byly použity dvě úlohy: (1) pasivní sledování 69 dvojic obrázků, z nichž každá zobrazovala kombinaci ultra-zpracovaných nebo minimálně zpracovaných potravin nebo ne-jídla, a (2) aktivní výběr preferované potraviny ze 23 dvojic obsahujících ultra- a minimálně zpracované potraviny. Viz Obr. 1.

Stimuly byly vybírány z databáze Food-pics (Blechert et al., 2014) a byly upraveny tak, aby byly vizuálně vyvážené z hlediska kontrastu, sytosti barev a velikosti. Při výběru podnětů jsme dále zohledňovali srovnatelnost kategorií z hlediska vizuální komplexity a luminance. Každá dvojice byla prezentována po dobu 3 sekund v první úloze, ve druhé pak po dobu trvání odpovědi participanta.

Prezentace probíhala na 19" LCD monitoru se šedým pozadím. Eye-trackingová data byla zaznamenávána pomocí systému EyeLink 1000 Plus s frekvencí 500 Hz.



Obr. 1: Podnětový materiál

2.3 Čištění dat a postup analýzy

Celkem bylo sesbíráno 98 datových záznamů. Po vícestupňovém čištění bylo vyloučeno 21 participantů. Sedm dětí bylo vyřazeno kvůli závažným zrakovým vadám (např. nekorigovaná dalekozrakost, barvoslepost, strabismus). Další čtyři pro nadměrnou ztrátu signálu, jeden kvůli špatné validaci (chyba >1°) a tři pro extrémně nízký počet fixací. Na základě laboratorních poznámek byli vyřazeni dva účastníci s poruchou příjmu potravy a čtyři s výraznými projevy ADHD. Do finální analýzy bylo zahrnuto 77 dětí.

AOI (oblasti zájmu) byly předem definovány jako obdélníkové oblasti kolem každé z potravin na obrazovce. Hlavními výstupy byly celková délka fixací na jednotlivé AOI a počet fixací. Pro analýzu byla použita kombinace deskriptivní statistiky, smíšených regresních modelů (LMM a GLMM) a analýza rozptylu (ANOVA) k porovnání mezi skupinami. V případě preferenční úlohy byla hlavní proměnnou binární volba jednoho ze dvou obrázků.

2.4 Etika výzkumu

Výzkum byl schválen Etickou komisí Mendelovy univerzity. Účast ve výzkumu byla dobrovolná. Před zařazením do studie byl získán písemný informovaný souhlas zákonných zástupců dětí a zároveň ústní souhlas samotných dětských participantů. Studie byla realizována v souladu s etickými principy výzkumu s lidskými účastníky kodifikovanými v Helsinské deklaraci (1964).

3 Výsledky

V této části jsou uvedené pouze částečné výsledky zaměřené na porovnání obou skupin, kompletní výsledky jsou zpracovávány do výzkumného článku.

3.1 Dotazník DEBQ-C

Analýza ukázala, že děti s obezitou měly signifikantně vyšší skóre v subškále restriktivního stravování (průměr = 2.22, SD = 0.80) než kontrolní skupina (průměr = 1.65, SD = 0.73), t(73.49) = -3.29, p = .002, d = -0.75.

3.2 Fixace při volném sledování

Pro porovnání dwell time (tedy suma fixací) na jednotlivých typech AOI u obou skupin byl využit smíšený lineární model. Pokud se zaměříme na within-subject plánované kontrasty, tak u ultra-zpracovaného jídla kontrolní skupina (průměr = 1220 ms) vykazovala statisticky významně delší dwell time než experimentální skupina (průměr = 1125 ms), t(Inf) = 4.34, p < .001, d = 0.17. Opačný vzorec byl detekován v případě minimálně zpracovaného jídla, kde vykazovala experimentální skupina (průměr = 1275 ms) delší dwell time, než kontrolní (průměr = 1210 ms), t(Inf) = -2.94, p = .003, d = -0.11. Stejný vzorec byl pozorován i podnětů zobrazujících ne-jídlo, kde byl dwell time u experimentální skupiny (průměr = 1185 ms) opět delší než u kontrolní (průměr = 1133 ms), t(Inf) = -2.37, p = .018, d = -0.09. Přestože uvedené kontrasty dosáhly statistické významnosti, velikosti efektu byly nízké až velmi nízké. Tyto výsledky je proto nutné interpretovat obezřetně: naznačují spíše slabý posun ve vzorci vizuální pozornosti než robustní či prakticky výrazný rozdíl mezi skupinami.

3.3 Preference

Na kalkulaci rozdílů v preferencích u druhé úlohy byl využit GLMM model s binomiální logit link funkcí. Výstupem tohoto modelu byla pravděpodobnost výběru ultra-zpracovaných potravin u obou skupin. Výsledky plánovaných kontrastů naznačují, že pravděpodobnost výběru ultra-zpracovaných potravin je u experimentální skupiny statisticky významně nižší (32.1 %) než u skupiny kontrolní (45.3 %), b = -0.564, p = .012.

3.4 DEBQ-C a experimentálních úloh

Nakonec byly vypočítány dva smíšené regresní modely se subškálami DEBQ-C jako prediktory a s dwell time (úloha 1) resp. preferencemi (úloha 2) jako závisle proměnnými. Žádná z dimenzí dotazníku významně nevysvětlovala dwell time u úlohy 1. U úlohy 2 byla dimenze restriktivního stravování v signifikantním vztahu s preferencí ultra-zpracovaného jídla, čím vyšší bylo skóre v této subškále, tím nižší byla pravděpodobnost výběru ultra-zpracovaného jídla ($b = -0.26$, $p = .022$).

4 Závěr

Zjištění této studie jsou v rozporu s očekáváním, že děti s obezitou vykazují zvýšenou vizuální citlivost vůči vysoce kalorickým podnětům. Celkově výsledky naznačují možný inverzní vzorec, tedy tendenci dětí s obezitou věnovat relativně méně pozornosti ultra-zpracovaným potravinám a častěji preferovat zdravější alternativy. Vzhledem k malým velikostem efektů je však třeba tato zjištění chápat jako předběžná a interpretovat je opatrně. Možným vysvětlením může být socializační vliv prostředí, ve kterém se děti nacházejí, včetně zdravotních a vzdělávacích intervencí během pobytu v lázních. Dalším faktorem může být zvýšená vnitřní motivace ke změně stravovacích návyků. Výsledky zároveň upozorňují na omezenou platnost teorie pozornostního zkreslení u dětí, u nichž může hrát větší roli kontrolní mechanismus a normy. Z praktického hlediska mohou tato zjištění předběžně naznačovat, že při dalším výzkumu behaviorálních intervencí zaměřených na stravovací volby dětí může být užitečné věnovat pozornost nejen omezení dostupnosti nezdravých potravin, ale také motivačním a normativním faktorům. Vzhledem k observační povaze prezentovaných výsledků a specifickému kontextu náboru je však nutné tyto implikace chápat jako hypotetické a vyžadující další ověření.

Poděkování

Tento výzkum byl financován Technologickou agenturou České republiky (TAČR Sigma, TQ01000275; *Chronic non-communicable disease prevention tools: innovative approaches and tools for assessing child health and obesity risks*). Děkujeme všem dětem a personálu dětských sanatorií za spolupráci.

Literatura

Blechert, J., Meule, A., Busch, N. A., & Ohla, K. (2014). Food-pics: An image database for experimental research on eating and appetite. Frontiers in Psychology, 5, 617. https://doi.org/10.3389/fpsyg.2014.00617

Graham, R., Hoover, A., Ceballos, N. A., & Komogortsev, O. (2011). Body mass index moderates gaze orienting biases and pupil diameter to high and low calorie food images. Appetite, 56(3), 577–586. https://doi.org/10.1016/j.appet.2011.01.033

van Strien, T., Oosterveld, P., & Reynaert, C. (2008). Children's DEBQ: Dutch Eating Behavior Questionnaire for Children. Appetite, 50(1), 56–62. https://doi.org/10.1016/j.appet.2007.05.006

Wang, Y., & Lobstein, T. (2006). Worldwide trends in childhood overweight and obesity. International Journal of Pediatric Obesity, 1(1), 11–25. https://doi.org/10.1080/17477160600586747

Werthmann, J., Jansen, A., Vreugdenhil, A. C. E., Nederkoorn, C., Schyns, G., & Roefs, A. (2015). Food through the child's eye: An eye-tracking study on attentional bias for food in healthy-weight children and children with obesity. Health Psychology, 34(12), 1123–1132. https://doi.org/10.1037/hea0000225

Sex s roboty: perverze, parafilie a etika umělé intimity

David Černý

Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 271/2, 182 07 Praha 8
Email: cerny@cs.cas.cz

Abstrakt

Studie analyzuje sexuální roboty z perspektivy sexuologie, filozofie sexu a etiky. V první části sleduje historický posun od teologického pojmu perverze přes medicínskou patologizaci k soudobé sexuologické kategorii parafilie a parafilní poruchy. Ve druhé části představuje hlavní filozofické teorie sexuální perverze – od Nagelovy fenomenologie reciproční touhy přes personalistické přístupy Solomona a Scrutona k deflačním koncepcím Goldmana a Primoratze – s otázkou robotického partnera. Těžiště studie tvoří etická analýza argumentů pro a proti sexuálním robotům se zaměřením na autonomii, objektifikaci, genderovou subordinaci, ochranu soukromí a komerční exploataci intimity. Závěr téma zasazuje do kontextu přechodu ke Společnosti 5.0 a klade otázku, jak budou kyberfyzické systémy proměňovat samotnou strukturu intimních vztahů.

1 Úvod: roboti, umělá těla a stará fascinace živým strojem

Sexuální roboti se mohou zdát jako okrajová součást současné robofilozofie, ve skutečnosti však vyvolávají řadu otázek, které jsou pro filozofii techniky, etiku i filozofii sexuality mimořádně zajímavé. Jednoduchá pracovní charakteristika říká, že jde o roboty, jejichž účelem je poskytovat člověku sexuální uspokojení. Takové vymezení je záměrně minimalistické: současné výrobky jsou většinou spíše hybridem realistických sexuálních panen, omezené mechaniky, senzoriky, hlasové interakce a nástrojů umělé inteligence. Lze dokonce pochybovat o tom, zda – striktně vzato – něco takového jako sexuální robot skutečně existuje. Filozofické a etické úvahy se ale nesoustředí pouze na současnost s jejím stavem vědecko-technického vývoje. Zaměřuje se také na to, co se stane, až stále dokonalejší umělé bytosti začnou intenzivněji vstupovat do rolí, které jsme tradičně vyhrazovali lidem: milencům, partnerům, důvěrníkům a objektům romantické lásky.

Z historického hlediska nepředstavují sexuální roboti nějaké překvapivé vybočení z evolučně podmíněných projevů lidské sexuality. Lidé si sexuální zkušenost dlouhodobě rozšiřují pomocí předmětů, obrazů, fantazií a umělých náhražek. Archeologické nálezy paleolitických falů, pozdější erotické pomůcky, *dame de voyage*, které zřejmě používali námořníci na dlouhých plavbách k uspokojení svých sexuálních potřeb, gumové a silikonové panny i současné robotické modely tvoří kontinuum, v němž se mění materiál, realističnost a míra interaktivity, nikoli základní tendence lidské sexuální imaginace (Ferguson, 2010; Listerová, 2022; Levy, 2008). Moderní vývoj začíná u realistických silikonových těl, pokračuje ranými modely inzerovanými jako sexuální roboti, jako je např. Roxxxy, a dospívá k současným umělým tvorům, které kombinují antropomorfní vzhled, jednoduchou mimiku, erotické zvuky a konverzační rozhraní. Pravé sexuální aktérství jim ale zatím chybí, současný sexuální robot je tedy zatím spíše příslibem než plně sexuálně funkčním umělým tvorem , ale i samotný příslib stačí k tomu, aby aktivoval naše sociální a sexuální skripty.

Fascinace živým nebo oživeným artefaktem je starší než robotika. Ovidiův Pygmalión touží po dokonalé ženě, kterou si sám vytvořil, a pozdější tradice z Galateie učinila paradigmatickou postavu umělé (z neživé hmoty stvořené) milenky. Adrienne Mayorová tento mýtus vykládá jako jeden z prvních příběhů o ženském androidovi, s nímž člověk vstupuje do vztahu i sexuálního spojení (Ovidius Naso, 2023; Mayor, 2019). Podobné motivy se objevují i jinde; u Villiersе de l'Isle-Adamovy Hadaly, Hoffmannovy Olimpie i v indických vyprávěních o mechanických bytostech. Tyto příběhy jsou podmanivé, ale zároveň znepokojivé: umělá žena je v nich často pasivní, tichá, projekční plocha mužské touhy. Dějiny sexuálních robotů proto od počátku nejsou jen dějinami techniky, nýbrž také dějinami fantazie, objektifikace, touhy a obavy, že umělá intimita může narušit ustálený řád lidských vztahů (Hoffmann, 2016; Cohen, 2002; Wosk, 2024).

2 Od perverze k parafilii

Představa sexu s robotem v nás může vyvolávat nepříjemné emoce, pocit, že překračujeme důležitou hranici, za níž bychom se vydávat neměli. I když je lidská sexualita pestrá, zřejmě vždy existovaly nějaké hranice – někdy dokonce velmi úzké a striktní – oddělující přijímané, povolené, „normální“ a perverzní. V křesťanské morální teologii, paradigmaticky u Tomáše Akvinského, byl sex posuzován z hlediska

přirozeného řádu: správný je tehdy, když odpovídá manželskému spojení muže a ženy a zůstává otevřený plození. Aktivity jako masturbace, sex se zvířaty nebo análním sex Tomáš chápal nejen jako *contra rationem*, ale i jako *contra naturam*, tedy nejen proti řádu rozumu a lidských institucí (např. nevěra je prohřešek proti instituci manželství), ale proti samotnému řádu přírody. Sex s robotem by v tomto morálně teologickém rámci musel být považovaný za zvlášť závažné překročení norem sexuální morálky: není součástí manželského svazku muže a ženy (*contra rationem*) a nemůže naplnit reprodukční účel sexuality (je *contra naturam*, nepřirozený).

Výraz *perverze* si dnes spojujeme prakticky výlučně se sexem. Latinské slovo *pervertere* má ale celou řadu významů: převrátit, zvrátit či zkazit; *perversus* je křivý, zvrácený, mravně pokřivený. V teologickém a právním kontextu proto perverze označovala narušení přirozeného nebo Bohem daného řádu, který se samozřejmě týkal i toho, co se chápalo jako přirozený řád lidské sexuality. Sexuální odchylka mohla být chápána jako hřích (teologická kategorie), zločin (právní kategorie) i společenská hrozba. Tresty za sexuální perverze, např. zoofilii, mohly být tresty nejpřísnějšími, jak dokládají zachovalé zapisy soudních případů, v nichž se například za sex se psem ukládal trest nejvyšší. K důležitému posunu došlo v 18. století s masturbační panikou vyvolanou spisem anonymního autora*Onania* a Tissotovým pojednáním o onanii *A Treatise on the Crime of Onan*. Masturbace se zde ještě morálně odsuzuje jako nepřirozená aktivita, zároveň se však stále silněji medikalizuje: má způsobovat slabost, impotenci, melancholii, poškození rozumu nebo dokonce smrt (Tissot, 1766). Zde se objevují důležité prvky charakterizující pozdější přechod od teologického, morálního a právního chápání sexuální perverze (tedy kategoide jednání, jež zasluhuje odsouzení a trest) k jejímu uchopení medicínskému jako specifické formy nemoci.

V 19. století se tento vývoj promítá do rodící se sexuologie. Heinrich Kaan, dnes již pozapomenutý zakladatel sexuologie, chápe perverze jako pokřivení pohlavního pudu, jehož přirozeným cílem je reprodukce. Mezi tyto odchylky (perverze) onanii, pederastii, lásku mezi ženami, nekrofilii, sex se zvířaty i sexuální zaujetí sochami (Kaan, 2016). Krafft-Ebingova *Psychopathia Sexualis*, zakládající dílo moderní sexuologie, dále rozvíjí model sexuální patologie: odchylky od heterosexuálního styku podřazenému evolučnímu cíli rozmnožování a zachování druhu jsou popisovány jako projev nemoci, degenerace nebo poruchy pohlavního pudu (von Krafft-Ebing, 2011). Původní paradigma perverze se soustředilo na hledání příčin perverze v degenerativních procesech lidského těla, s čímž se o několik desítek let později radikálně rozešel zakladatel psychoanalýzy Sigmund Freud ve svém díle *Tři pojednání o teorii lidské sexuality*. Freud sice tento biologizující rámec zásadně proměnil, protože lidskou sexuality nepodřazoval nějakému biologickému cíli (reprodukci a zachování druhu), přesto si ale zachoval ideál zralé sexuality za kměřené výlučně na vaginálnímu styk muže a ženy. Freud by sex s robotem nejspíše popsal jako fetišistickou náhradu: sexuální zaměření na objekt, který zastupuje lidské tělo, člověkem ale není (Freud, 2016).

Moderní sexuologie se od pojmu perverze z dobrých důvodů odklonila. Hovoří spíše o parafilii a parafilní poruše. Parafilie označuje přetrvávající sexuální zájem odchylující se od běžného souhlasného sexu mezi dospělými, sama o sobě ještě není poruchou. Parafilní poruchou se stává teprve tehdy, když působí utrpení, narušení života nebo újmu druhým. Toto rozlišení je pro sex s roboty klíčové. V současné kultuře může jít o parafilii, protože sexuální zájem o robotické tělo není obvyklý. Z toho však neplyne, že jde o patologii, něco, čím by se měli zabývat lékaři a společnost znepokojovat. Rozhodující není fakt, že partnerem je stroj, nýbrž praktické a psychické důsledky: zda sex s umělou bytostí praxe škodí uživateli, druhým lidem nebo sociálním vztahům. Pokud nepůsobí újmu, nelze ho jednoduše označit za poruchu jen proto, že se nám zdá nezvyklý, nepřirozený nebo esteticky odpudivý (Hock, 2012; Hyde a DeLamater, 2024). Z tohoto moderního výkladu parafilie plyne, že již dnes v některých společnostech (a zřejmě v budoucnosti i jinde) sex s robotem zřejmě parafilií není (protože se nevnímá jako podivný či neobvyklý). A současný vývoj možná směřuje ke společnosti, který bude sexuálně otevřenější, tolerantnější a také mnohem pestřejší.

3 Perverze ve filozofii sexu a sex s roboty

Filozofie sexu se k pojmu perverze vrací jinou cestou než sexuologie. Neptá se primárně po diagnóze, nýbrž po povaze sexuální touhy, sexuálního aktu a hodnot, které v sexu nacházíme. Thomas Nagel ve vlivném článku *Sexual Perversion* chápe dobrý sex jako komplex vzájemně se zrcadlících tužeb: lidé po sobě touží, vnímají své touhy a chtějí být objektem této touhy. Perverze je porušením této reciproční struktury (Nagel, 1969). Pokud tento model aplikujeme na sex s robotem, problém je zřejmý: robot po člověku netouží a nemůže prožívat pocity a prožitky, které zakouší člověk. Sex s robotem by tak byl neúplným sexem, perverzí, sexem založeným na iluzi reciprocity.

Další filozofické koncepce sexu na něj kladou podobně náročné požadavky. Robert Solomon spojuje sex s širší strukturou uznání a mezilidského vztahu, Roger Scruton jej chápe jako intencionální vztah k vtělené osobě přirozeně se vyvíjející směrem k lásce (Solomon, 1975; Scruton, 1986). Tyto teorie sexu kladou podmínky, které robot splňuje jen obtížně, protože mu chybí vědomí, prožitky, zkušenosti, touha i schopnost skutečné reciprocity. Alan Goldman a Igor Primoratz však nabízejí minimalističtější (a reálnější)

analýzu sexu: je to aktivita zaměřená na tělesnou slast, jejíž morální hodnota závisí především na souhlasu, neškození a kontextu, nikoli na nutném spojení s láskou, reprodukcí nebo hlubokým meziosobním vztahem (Goldman, 1977; Primoratz, 1999). Z této perspektivy nemusí být sex s robotem perverzí, špatným či méně hodnotným sexem. Může být neobvyklý, zvláštní nebo kulturně provokativní, ale není proto automaticky perverzí, špatným či nemorálním sexem.

Měli bychom rozlišit dvě otázky. První zní, zda sex s robotem splňuje nějaké ideály erotické zkušenosti dvou lidských osob. Zde je odpověď nejspíše záporná, alespoň u současných robotů. To ale ze sexu s roboty automaticky nedělá špatný sex, navíc zůstává otevřenou možnost, že roboti budoucnosti budou schopni plnohodnotného sexu. Druhá otázka zní, zda je sex s roboty nemorální nebo patologický. A odpověď je záporná. Lidská sexualita není jednorozměrná; zahrnuje reprodukci, slast, hru, intimitu, úlevu, experiment, terapii i fantazii. Sexuální robot může v některých případech selhávat jako partner v silném personalistickém smyslu, ale přesto může fungovat jako erotický artefakt, s nímž uživatel vědomě vstupuje do imaginativní, tělesné a afektivní hry. Zde se uplatňují psychologické mechanismy umožňující iluzi sexuální interakce: robot nemusí skutečně toužit a prožívat, aby v nás spustil mechanismy antropomorfizace, projekce, imerze a sexuálního vzrušení (Szczuka a spol., 2019).

4 Etika sexu s roboty: argumenty pro a proti

Etická analýza sexu s roboty by neměla začínat možnými pocity znechucení, které v lidech může vyvolávat. Znechucení a intuitivní pocity odporu jsou psychologicky pochopitelné, ale normativně nespolehlivá. Dějiny lidské sexuality ukazují, že mnohé praktiky byly odsuzovány jako nepřirozené, dokud se neproměnil širší kulturní rámec jejich interpretace. To neznamená, že je vše dovoleno. Znamená to pouze, že kritériem správnosti a nesprávnosti nemůže být pouhá odchylka od tradice, zvyklostí, kulturních norem a očekávání. V případě sexu s roboty je třeba etickou analýzu zaměřit především autonomii, újmu, dopady na vztahy, sociální spravedlnost, objektifikaci, ochranu soukromí, osobních a citlivých dat.

Argumenty ve prospěch sexuálních robotů se opírají například o rozšíření sexuální autonomie. Pokud dospělý člověk používá robotické tělo k sexuálnímu uspokojení, rozšiřuje si možnosti sexuálních zkušeností a obohacuje svůj sexuální život. Jeho autonomní volby samy o sobě nepředstavují zásah do práv druhých a nejsou zdrojem újmy. Sexuální roboti mohou také pomoci lidem, kteří mají z různých důvodů omezený přístup k sexuálním vztahům: seniorům, lidem s postižením, osamělým osobám, lidem s nějakým traumatem nebo těm, kteří se obtížně vyrovnávají se studem, vzhledem vlastního těla či třeba trpí poruchami spojenými s prožívaným tlakem na sexuální výkonnost (Di Nucci, 2017; Jecker, 2021; McArthur, 2017). Sex navíc má výrazné zdravotní a psychologické benefity: slast, uvolnění, posílení sebevědomí, redukci stresu a obnovení pocitu sexuálního aktérství (Neves, 2023). Robot nemusí nutně nahrazovat lidského partnera; může fungovat jako pomůcka, přechodná opora nebo specifický nástroj erotického sebevzdělávání.

Některé argumenty ve prospěch (nějaké formy) využívání sexuálních robotů se soustředí na snižování možné újmy (*harm-reduction*). Uvažuje se o tom, zda by sexuální roboti mohli snižovat rizikové nebo násilné chování tím, že nabídnou bezpečný prostor pro realizaci neobvyklých a rizikových sexuálních fantazií nebo poskytnou přístup k sexu mladým a frustrovaným mužům (incelové), kteří se dopouštějí násilí na ženách. Tyto argumenty je však třeba formulovat opatrně. Zatím není dostatečně empiricky doloženo, že by využívání robotů snižovalo riziko újmy. Přesto nelze předem vyloučit, že minimálně v kontrolovaném terapeutickém kontextu mohou roboti u některých osob přispět k tomu, že se nebudou dopouštět jednání, které ubližuje druhým. Z etického hlediska je důležité vyvažovat negativa a pozitiva sexu s roboty nejen v individuální, ale i širší společenské rovině. Například jednání incelů je pochopitelně zavrženíhodné. Pokud by ale sex s roboty vedl k tomu, že se mladí muži budou méně radikalizovat a omezí se jejich nenávistné projevy vůči ženám, můžeme to chápat jako významný společenský přínos, která si zaslouží naší podporu.

Nejzávažnější námitky proti využívání sexuálních robotů se týkají objektifikace a genderových norem. Kathleen Richardsonová a další kritičky tvrdí, že sexuální roboti reprodukují logiku pornografie, prostituce a ženské podřízenosti: představují ženská těla určená k použití, bez autonomie, vlastních zájmů a nároků (Richardson, 2022; Roper, 2022). Tato námitka je zřejmě nejsilnější v případech, kde design robotů skutečně stylizuje ženy jako pasivní, poddajné, objekty, s nimiž si uživatel může dělat vše, co se mu zamane. Všeobecný zákaz využívání sexuálních robotů však z této námitky neplyne. Objektifikace je vlastností sociální praxe, reprezentce a vztahu uživatelů robotů k druhým lidem; nevyvstává automaticky ze samotné existence sexuálních robotů. Sexuální roboti mohou mít řadu podob a vlastností a nemusí být navrhováni jako poslušné náhražky žen. Dokonce ani nemusí mít lidskou či plně lidskou podobu (Devlin, 2018, 2020). Kromě toho, postoje lidí k robotům jsou pestřejší než by tyto námitky připouštěly a jedním ze silných mechanismů určujících naše vztahy k robotům je zlidšťování, které lze chápat jako opak objektifikace.

Další skupina námitek se týká soukromí, dat a závislosti. Sexuální robot vybavený senzory, kamerami, mikrofony a konverzační umělou inteligencí

by mohl sbírat mimořádně citlivá data o těle, preferencích, fantaziích, zranitelnosti i emocionálních projevech uživatele. V prostředí kapitalismu dohledu nejde o okrajové riziko, ale o strukturální problém: intimita se může stát datovou surovinou a vztah k robotovi prostředkem behaviorální predikce a monetizace (Zuboff, 2019; Brooks, 2021). Etická přijatelnost sexuálních robotů proto závisí nejen na tom, co dělá uživatel, ale také na institucionálním rámci: ochraně dat, transparentnosti, auditovatelnosti, zákazu manipulace a vytváření zranitelností, a možnosti skutečné kontroly nad zařízením.

Etická analýza sexu s roboty je komplexní. Sex s robotem není nemorální jen proto, že je neobvyklý, neumožňuje reprodukci a nahrazuje živé tělo člověka umělým tělem robotickým. Morálně relevantními faktory, které mohou nachýlit tuto analýzu v neprospěch sexuálních robotů, zahrnují újmu, negativní dopady na společnost, objektifikaci, oslabení autonomie, patologickou izolaci nebo prolamování soukromí a zneužívání osobních dat. Podaří-liu se tyto faktory eliminovat/minimalizovat, mohou pozitiva sexu s roboty převážit a sen se se může stát morálně neproblematickou součástí širší erotické kultury, v níž se lidská sexualita nevyčerpává reprodukcí ani ideály reciprocity či lidské výlučnosti.

5 Technika, intimita a Společnost 5.0

Téma sexu s roboty je třeba zasadit do širšího dějinného rámce. Lidská společnost byla vždy spoluutvářena technikou. Lovecko-sběračské skupiny žily v malých, relativně rovnostářských celcích, kde přežití skupiny určovalo podobu rolí, péče i partnerských vazeb. Objev zemědělství vedl ke vzniku multigeneračních domácností, hromadění vlastnictví a důrazu na dědění; intimní vztahy se silněji svázaly s rodem, majetkem a manželství bylo primárně ekonomicko-právní institucí. Průmyslová revoluce přesunula lidi do měst, oddělila práci od domova a podpořila vznikl nukleární rodiny, poprvé v dějinách lidstva se romantická láska stala uznávaným důvodem vzniku manželství. Informační revoluce z jednotlivce učinila uzel v sítích, v nichž se opora, identita a intimita částečně přesouvají mimo lokální komunitu i rodinu.

Nyní stojíme na přechodu do Společnosti 5.0, v níž se umělá inteligence, robotika, kyberfyzické systémy, biotechnologie a datové infrastruktury stanou nejen nástroji, ale aktivními účastníky sociálních procesů. Roboti budou vstupovat do péče, vzdělávání, medicíny, domácností i intimních vztahů. Stojíme na prahu revolučních změn: nejde už jen o to, jak technika nepřímo proměňuje společnost, nýbrž o to, jak budou technické systémy zaujímat role, které byly považovány za výlučně lidské. Budou nás léčit, opečovávat, doprovázet, utěšovat, naslouchat nám – a v některých případech s námi mít sex.

Závěr této práce není technofobní, ale ani naivně technoptimistický. Sexuální roboti mohou být výrazem nové svobody i nových forem závislosti; mohou rozšířit sexuální aktérství, ale také reprodukovat staré modely objektifikace; mohou pomoci osamělým, ale také prohloubit komercializaci intimity. Jisté je pouze to, že stojíme na prahu proměny, která zasáhne samotnou strukturu lidskách vztahů. Pokud předchozí technologické revoluce změnily rodinu, práci, domov a hranici mezi veřejným a soukromým, Společnost 5.0 pravděpodobně změní i to, koho budeme považovat za možného partnera vstupujícího do sfér naší intimity, včetně toho, s kým budeme mít sex.

Poděkování

Tento příspěvek vznikl za podpory programu AV 21 Epicentra civilizace – inteligentní domácnosti, technologie, společnost.

Reference

Brooks, R. (2021). *Artificial Intimacy: Virtual Friends, Digital Lovers and Algorithmic Matchmakers*. Columbia University Press, New York.

Cohen, S. (2002). Romancing the robot and other tales of mechanical beings in Ancient Indian literature. *Acta Orientalia*, 64:65–75.

Danaher, J. a McArthur, N. (eds.) (2017). *Robot Sex: Social and Ethical Implications*. The MIT Press, Cambridge, Mass.

Devlin, K. (2018). *Turned On: Science, Sex and Robots*. Bloomsbury, London.

Devlin, K. (2020). The ethics of the artificial lover. Liao, S. M. (ed.), V *Ethics of Artificial Intelligence*, s. 271–290. Oxford University Press, New York.

Di Nucci, E. (2017). Sex robots and the rights of the disabled. Danaher, J. a McArthur, N. (eds.), V *Robot Sex: Social and Ethical Implications*, s. 74–88. The MIT Press, Cambridge, Mass.

Ferguson, A. (2010). *The Sex Doll: A History*. McFarland & Company, Jefferson.

Freud, S. (2016). *Three Essays on the Theory of Sexuality*. Verso, London.

Goldman, A. (1977). Plain sex. *Philosophy and Public Affairs*, 6(3):267–287.

Hock, R. R. (2012). *Human Sexuality*. Pearson, Boston, 3rd. vyd.

Hoffmann, E. T. A. (2016). *Noční kousky*. Volvox Globator, Praha.

Hyde, J. S. a DeLamater, J. D. (2024). *Understanding Human Sexuality*. McGraw Hill, New York, 15th. vyd.

Jecker, N. S. (2021). Nothing to be ashamed of: Sex Robots for older adults with disabilities. *Journal of Medical Ethics*, 47:26–32.

Kaan, H. (2016). *Psychopathia Sexualis*. Cornell University Press, Ithaca.

Levy, D. (2008). *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. Harper, New York.

Listerová, K. (2022). *Podivuhodná historie sexu*. Paseka, Praha.

Mayor, A. (2019). *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton University Press, Princeton.

McArthur, N. (2017). The case for sexbots. Danaher, J. a McArthur, N. (eds.), V *Robot Sex: Social and Ethical Implications*, s. 31–45. The MIT Press, Cambridge, Mass.

Nagel, T. (1969). Sexual perversion. *Journal of Philosophy*, 66(1):5–17.

Neves, S. (2023). *Sexology: The Basics*. Routledge, London.

Ovidius Naso, P. (2023). *Proměny*. Městská knihovna v Praze, Praha.

Primoratz, I. (1999). *Ethics and Sex*. Routledge, London.

Richardson, K. (2022). The end of sex robots: Porn robots and representational technologies of women and girls. Richardson, K. a Odlind, C. (eds.), V *Man-Made Women*, s. 171–192. Palgrave Macmillan, Cham.

Roper, C. (2022). *Sex Dolls, Robots, and Woman Hating: The Case for Resistance*. Spinifex Press, North Geelong.

Scruton, R. (1986). *Sexual Desire: A Philosophical Investigation*. Weidenfeld & Nicolson, London.

Solomon, R. (1975). Sex and perversion. Baker, R. a Elliston, F. (eds.), V *Philosophy & Sex*, s. 268–287. Prometheus Books, Buffalo, NY.

Szczuka, J. M., Hartmann, T. a Krämer, N. C. (2019). Negative and positive influences on the sensations evoked by artificial sex partners. Zhou, Y. a Fischer, M. H. (eds.), V *AI Love You*, s. 3–19. Springer, Cham.

Tissot, S. A. D. (1766). *A Treatise on the Crime of Onan*. Strand, London, 3rd. vyd.

von Krafft-Ebing, R. (2011). *Psychopathia Sexualis*. Arcade Publishing, New York.

Wosk, J. (2024). *Artificial Women: Sex Dolls, Robot Caregivers, and More Facsimile Females*. Indiana University Press, Bloomington.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books, London.

Constructing Grounded Conceptual Representations via Object-Centric Active Vision

Miroslav Cibula, Kristína Malinovská, Igor Farkaš

Faculty of Mathematics, Physics and Informatics
Comenius University Bratislava
cibula25@uniba.sk, kristina.malinovska@fmph.uniba.sk, igor.farkas@fmph.uniba.sk

Abstract

High-level human cognition depends on structured representations enabling entities, attributes, and events to be compared, categorized, and reused across contexts. Raw perceptual input, however, is continuous and non-symbolic, and to achieve a higher level of conceptual understanding, a cascade of transformations must be performed on it to translate it effectively. The present work aims to explore how visual perceptual information might be computationally organized into structured conceptual representations that can support concept formation and basic reasoning. We propose a preliminary active computer vision system that processes a 3D video stream to segment scenes, track objects, infer simple object properties, and represent affordances through interaction. Via this process, the architecture is intended to construct object-centric, perceptually grounded conceptual representations within the framework of conceptual spaces, embedding observed object attributes and motion behaviors into a structured semantic space. This approach aims to draw on cognitively inspired principles, prioritizing modularity and interpretability over state-of-the-art fully monolithic, end-to-end neural architectures. Consequently, this perceptual architecture could complement other reasoning and learning systems in the future, contributing to a more robust cognitive architecture applicable in the real world.

1 Introduction and Related Work

Visual perceptual input is an essential modality for early sensorimotor discovery and learning in human infants (Gopnik et al., 1999). Especially during pre-verbal developmental stages, visual perception, coupled with the ability to physically interact with the environment, is theorized to facilitate the construction of multimodally grounded semantic representations (Fitzpatrick et al., 2003; Lohmann et al., 2020). The semantics of early conrete concepts in this stage are not primarily learned via language (Mandler, 1992; Carey, 2009), but rather by observing the attributes (i.e., in the case of concrete physical objects, these attributes could be related to appearance, physical properties, behavioral patterns under some physical intervention by the agent, etc.) and characterizing the concepts using the sets of such attributes (Rosch, 1978; Gärdenfors, 2000; Takáč, 2007; Baillargeon, 2004; Held and Hein, 1963).

The theory of conceptual spaces (Gärdenfors, 2000, 2014) is one of the well-established computational frameworks modeling such semantic mechanisms with a higher degree of biological plausibility. Within it, observed instances of concepts, exemplars (Medin and Schaffer, 1978; Nosofsky, 1984), are aligned with other exemplars in the semantic space based on the similarity of their attributes. Consequently, the observed exemplars tend to cluster along the semantic space dimensions corresponding to their common attributes (i.e., quality dimensions), forming convex regions that represent conceptual categories, with the cluster centroids approximating prototypical instances (Rosch, 1978). As a result, conceptual spaces are formally hyperdimensional geometric spaces organized along interpretable and separable quality dimensions, supporting semantic similarity assessment by measuring spatial distances between exemplars or categories, as well as categorization and category recall, since any newly observed exemplar can be mapped to the appropriate convex region.

Hence, such a combination of a computational visual processing pipeline with conceptual spaces for semantic representations and an embodied agent leveraging physical manipulation to elicit new environmental observations may yield an approximate computational model of a bottom-up visual perception process (McMains and Kastner, 2011), specifically producing non-verbal physical-object-centric concepts grounded in perception and action (Barsalou, 1999). In the context of cognitive robotics, this process is essential for constructing high-level symbolic representations of the environment, facilitating higher-level reasoning and planning.

The integration of robotic interaction with computer vision methods to boost their performance, however, has been explored relatively extensively. Most notably, the research has focused on learning the properties of physical objects in the scene via motion tracking following the physical interaction performed by an embodied agent (Fitzpatrick et al., 2003; Fitzpatrick, 2003; Katz et al., 2013; Nematollahi et al., 2020; Kruzliak et al., 2024). Additionally, there has been a focus on learning manipulation strategies to explore the envi-

ronment more efficiently (Lohmann et al., 2020; Gadre et al., 2021), or on linking robotic actions to object perception to construct affordance representations (Gupta and Davis, 2007; Gadre et al., 2021).

The present paper proposes an artificial perceptual architecture (Section 2) that would leverage active computer vision, intended to enable robotic agents to study and learn the attributes and behaviors of physical objects and to generalize across multiple observations to form structured conceptual representations. In our design, we rely on standard computer vision and machine learning methods to facilitate this kind of visual processing and assemble them into a modular system functionally inspired by pertinent aspects of the human visual system. Thus, the proposed architecture is intended to remain transparent and interpretable, while preserving a cognitively inspired organization. Subsequently, we suggest how the architecture and the representations it yields could provide a substrate for more complex cognitive functions, such as language processing and primitive social cognition (Section 3). Our design methodology is hence in accordance with the developmental perspective of human cognition: a child first learns foundational, object-centric representations through sensorimotor exploration of the physical world (Mandler, 2007), subsequently grounds early linguistic labels in these pre-existing perceptual concepts (Tomasello, 2003; Bloom, 2000), and eventually leverages this acquired linguistic system as a cognitive tool to bootstrap the formation of increasingly abstract and complex representations (Vygotskij, 1978; Lupyan, 2012).

2 Proposed Architecture

The minimal computational high-level formulation of the proposed perceptual architecture could be defined as a stateful ϑ-parametrized function $\Pi_\vartheta \colon \boldsymbol{v}(t) \mapsto \boldsymbol{o}(t)$, where $\boldsymbol{v}(t) \in [0.0; 1.0]^{H \times W \times 4 \times c}$ denotes a visual observation at the discrete environmental time $t \in \mathbb{N}$, in our context defined as a sequence of $1 \leq c \leq t$ normalized RGB-D images, and $\boldsymbol{o}(t)$ being an abstract set of high-level disentangled state variables [akin to the ones in our previous work (Cibula et al., 2024)] extracted from the visual percept $\boldsymbol{v}(t)$.

The parametrization ϑ affects the system's subprocesses by specifying configurable hyperparameters of the modules within the architecture, with specifics depending on the system implementation. In our conceptualization, ϑ enables top-down influence from the rest of the architecture, thereby allowing a computational analog of top-down/bottom-up bidirectionality in the perceptual process (Dijkstra et al., 2017; McMains and Kastner, 2011). More specifically, ϑ can be interpreted as serving two purposes: capturing longer-term perceptual adaptation based on feedback from a potential reasoning component, and providing shorter-term control, for example, in attentional filtering (Deutsch and Deutsch, 1963; Cowan, 1998).

In designing the internal mechanisms of Π_ϑ (Figure 1), we do not rely on monolithic, end-to-end neural architectures. Such models tend to be opaque and resistant to integration into larger, modular cognitive architectures in which specific perceptual domains must be isolated and manipulated. Instead, the proposed framework leverages a decoupled, multi-branch pipeline utilizing specialized computer vision methods. This design is functionally inspired by aspects of the human visual system, partitioning continuous visual perception into discrete object representations, kinematic tracking, and spatial topology extraction.

2.1 Scene Discretization and Visual Memory

As raw perceptual input is predominantly continuous and non-symbolic, a bottom-up quantization mechanism is useful for converting the continuous RGB-D signal into discrete entities. In human infants, object individuation is strongly supported by cues such as motion and physical cohesion (Spelke, 1990). Computationally, we propose to operationalize this using standard prompt-free instance segmentation models coupled with multi-object tracking.

To support object identity across changing viewpoints, brief occlusions, and manipulation, the system employs Ephemeral Eidetic Memory Storage (EEMS). As objects are tracked, the EEMS captures geometric and visual crops of the objects from various viewpoints, provided the observations meet heuristic quality thresholds, such as low visual similarity to already recorded observations or high image sharpness. EEMS serves as a functional analog to short- to mid-term visual memory, accumulating multiple object-centered observations of the physical entities in the scene.

2.2 Feature Extraction

Following scene discretization into tracked entities, the architecture processes the EEMS data through three parallel branches, extracting representations that map to fundamental domains of human object cognition: intrinsic visual appearance (the "what"), spatial topology (the "where"), and dynamic behavior (the "how"). This categorization is inspired by the two-stream processing hypothesis (Goodale and Milner, 1992), and the theory of affordances (Gibson, 2014). By isolating these modalities, the system generates separable representations ready for conceptual embedding.

Visual Domain To capture the "what" attributes of an object (e.g., color, shape, texture), the system avoids language-dependent models in favor of unsupervised representation learning. Relatively recently, a new class of slot-attention-based models (Locatello et al., 2020) has emerged that can disentangle raw pixel inputs into distinct, continuous latent factors (Stammer et al., 2024). When applied to sets of visual observations for each object stored in the EEMS, the architecture is in-

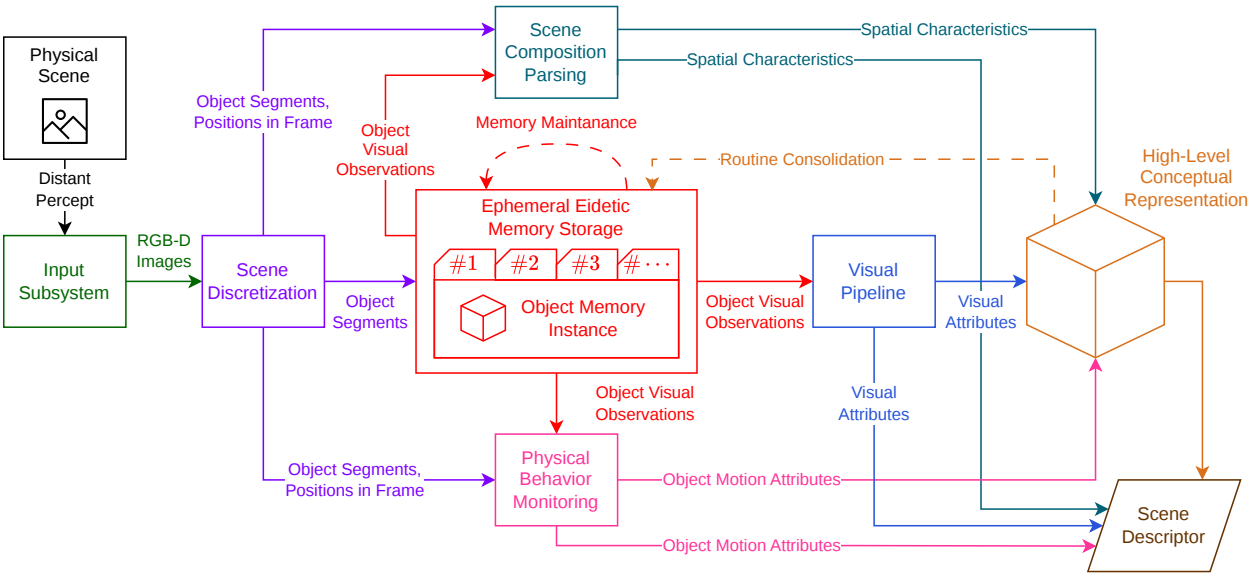


Fig. 1: High-level schematic of the proposed perceptual architecture with the essential subsystems facilitating the conversion of a low-level visual percept from a physical scene to a high-level scene description while learning high-level conceptual representations modeled as a conceptual space.

tended to approximate a decomposition of visual object properties into separable feature dimensions without requiring linguistic supervision.

Spatial Domain To represent the spatial organization of the scene, spatial relationships between objects can be formalized using purely geometric metrics, echoing developmental evidence that preverbal infants are sensitive to primitives such as spatial containment and support before acquiring spatial vocabulary (Mandler, 1992; Baillargeon, 2004). For each object, an egocentric spatial signature could be constructed by evaluating relative 3D directional vectors (i.e., azimuth, elevation, distance) and volumetric overlaps with surrounding objects. This approach would embed variable-sized scene graphs into fixed-length relational vectors, forming an embeddable spatial domain.

Kinematic Domain Finally, to characterize how objects behave and interact dynamically, the system can track 6-degree-of-freedom (6-DoF) trajectories of their movement. This could be achieved by any standard 6D object-pose estimator tracking each object during the recording session. Such a tracking process yields a 6D time-series trajectory that can be converted into compact motion descriptors. Acting as a signal, each trajectory can be filtered and transformed into the frequency domain (Mao et al., 2020). By retaining only the low-frequency coefficients, we can generate a compact, fixed-length vector that captures the prototypical "shape" of the motion (e.g., tumbling, sliding) and is less sensitive to the event's specific speed or duration.

2.3 Semantic Representation Construction

The objective of the perceptual architecture is to construct a structured semantic representation utilizing the framework of conceptual spaces (Gärdenfors, 2000).

The vectors extracted from the three processing branches (Section 2.2) would act as points within their respective, separable quality dimensions. Each observation of an object could function as an exemplar. As the embodied agent interacts with the environment, exemplars sharing common attributes can cluster together. The centroids of these clusters would emerge as the prototypical instances of concepts.

By measuring semantic similarity via spatial distances, the system could dynamically categorize novel objects by mapping them to the nearest established regions. This architecture would contribute to grounding conceptual meaning in multimodal perception and physical action (Barsalou, 1999; Harnad, 1990), providing a foundation for higher-level reasoning without requiring prior linguistic knowledge.

Concurrently, this categorization process yields the immediate high-level scene descriptor $\boldsymbol{o}(t)$. At any discrete time step t, $\boldsymbol{o}(t)$ serves as an instantaneous, object-centric state representation of the environment, encapsulating the tracked entities, their current kinematic information, spatial relationships, and their mapped conceptual categories. Consequently, the perceptual process yields a dual product: it progressively builds a longer-term, generalized semantic space while simultaneously producing a structured, real-time descriptor ready to be leveraged by downstream action-selection or reasoning modules.

3 Potential Cognitive Extensions

The proposed architecture suggests possible extensions toward higher-order cognitive capabilities while preserving the same object-centric representational princi-

ples. By producing a geometrically grounded semantic space, we hypothesize that the system could provide a useful representational substrate for language acquisition and primitive social cognition.

Language Grounding Following the framework of conceptual spaces, the proposed system represents object-related meaning geometrically prior to any linguistic involvement. Within this paradigm, linguistic labels can be modeled as pointers that guide attention to specific regions or centroids in conceptual space (Gärdenfors, 2024). This suggests a possible mapping between mathematical clusters and grammatical categories: visual attribute clusters may correspond to adjectives (e.g., "yellow", "cubic"), kinematic clusters may correspond to verbs (e.g., "slides", "falls"), and spatial clusters may correspond to prepositions (e.g., "behind", "on"). Consequently, attaching linguistic labels to these autonomously formed categories could support later compositional language inherently grounded in sensorimotor perception, thereby contributing to a possible solution to the symbol grounding problem (Harnad, 1990; Vogt, 2002).

Social Cognition Furthermore, the object-centric nature of the architecture suggests a possible pathway toward primitive social perception. By treating other agents as physical systems, the discretization module could segment a person into constituent body parts and track them as object-like entities. In such an extension, the system could implicitly approximate a form of body-pose representation, thereby extracting the behavioral kinematics and spatial topology of the observed agent. If the robotic agent has a model of its own body, it could measure semantic similarities between the tracked trajectories of the observed agent's limbs and its own end-effectors. This mapping between the self and the other could provide one prerequisite for observational imitation learning and simple intent recognition.

4 Conclusion

Here, we propose a theoretical framework for a modular active computer vision architecture designed to transform continuous non-symbolic video streams into discrete, high-level semantic representations. By explicitly avoiding opaque, end-to-end monolithic models, we outline a transparent, cognitively inspired pipeline that partitions visual processing into visual, spatial, and kinematic domains, drawing inspiration from human visual perception. By embedding these extracted features, we outline how the system could dynamically construct a conceptual space that reconciles exemplar-based observations with prototype-based category formation through geometric proximity, without linguistic influence.

Ultimately, this perceptual architecture is designed to serve as a non-verbal sensory module for embodied agents. By outputting real-time structural scene descriptors and progressively learned conceptual representations, the proposed system could provide the necessary grounding for downstream cognitive architectures, facilitating sensorimotor learning and inference, and potentially supporting later language grounding in embodied settings.

Acknowledgement

The authors thank Markus Vincze (Automation and Control Institute, TU Wien) for his advice and feedback. This research was supported by project APVV-21-0105. Research results were in part obtained using the computational resources of the supercomputer PERUN at the Supercomputing Center at TU Košice, with support from the EU, funds of the Recovery and Resilience Plan of the Slovak Republic, project 17I03-04-P03-00001. We also thank the Slovak Society for Cognitive Science (SSKV) for their support.

References

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, 13(3):89–94.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Learning, Development, and Conceptual Change. The MIT Press, Cambridge, MA, US.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Cibula, M., Kerzel, M., and Farkaš, I. (2024). Learning low-level causal relations using a simulated robotic arm. In *Artificial Neural Networks and Machine Learning*, pages 285–298. Springer Nature.

Cowan, N. (1998). Attentional filtering and orienting. In *Attention and Memory: An Integrated Framework*, pages 137–166. Oxford University Press.

Deutsch, J. A. and Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 70(1):80–90.

Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M. A. J., and Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, 7(1).

Fitzpatrick, P. (2003). First contact: An active vision approach to segmentation. In *Int. Conf. on Intelligent Robots and Systems*, pages 2161–2166. IEEE.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). Learning about objects through action - initial steps towards artificial cognition. In *Int. Conf. on Robotics and Automation*, pages 3140–3145. IEEE.

Gadre, S. Y., Ehsani, K., and Song, S. (2021). Act the part: Learning interaction strategies for articulated object part discovery. In *Int. Conf. on Computer Vision*, pages 15732–15741. IEEE.

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, NY.

Goodale, M. A. and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25.

Gopnik, A., Meltzoff, A., and Kuhl, P. (1999). *The Scientist in the Crib*. William Morrow and Company, Inc., New York, first Perennial edition.

Gupta, A. and Davis, L. S. (2007). Objects in action: An approach for combining action understanding and object perception. In *Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, Mass.

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Gärdenfors, P. (2024). Event structure, force dynamics and verb semantics. *Language Sciences*, 102:101610.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346.

Held, R. and Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5):872–876.

Katz, D., Kazemi, M., Andrew Bagnell, J., and Stentz, A. (2013). Interactive segmentation, tracking, and kinematic modeling of unknown 3D articulated objects. In *Int. Conf. on Robotics and Automation*, pages 5003–5010. IEEE.

Kruzliak, A. et al. (2024). Interactive learning of physical object properties through robot manipulation and database of object measurements. In *Int. Conf. on Intelligent Robots and Systems*, pages 7596–7603. IEEE.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T. (2020). Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538.

Lohmann, M., Salvador, J., Kembhavi, A., and Mottaghi, R. (2020). Learning about objects by learning to interact with them. In *Advances in Neural Information Processing Systems*, pages 3930–3941.

Lupyan, G. (2012). What do words do? Toward a theory of language-augmented thought. In Ross, B. H., editor, *The Psychology of Learning and Motivation*, volume 57, pages 255–297. Academic Press.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, 99(4):587–604.

Mandler, J. M. (2007). *Foundations of Mind: Origins of Conceptual Thought*. Oxford University Press.

Mao, W., Liu, M., and Salzmann, M. (2020). History repeats itself: Human motion prediction via motion attention. In *Computer Vision – ECCV 2020*, pages 474–489, Cham. Springer International Publishing.

McMains, S. and Kastner, S. (2011). Interactions of top-down and bottom-up mechanisms in human visual cortex. *Journal of Neuroscience*, 31(2):587–597.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Nematollahi, I., Mees, O., Hermann, L., and Burgard, W. (2020). Hindsight for foresight: Unsupervised structured dynamics models from physical interaction. In *Int. Conf. on Intelligent Robots and Systems*, pages 5319–5326. IEEE.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):104–114.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum, Hillsdale, N.J.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14(1):29–56.

Stammer, W., Wüst, A., Steinmann, D., and Kersting, K. (2024). Neural concept binder. In *Advances in Neural Information Processing Systems*, volume 37, pages 71792–71830. Curran Associates, Inc.

Takáč, M. (2007). *Construction of Meanings in Living and Artificial Agents*. PhD thesis, Comenius University Bratislava.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Vogt, P. (2002). The physical symbol grounding problem. *Cognitive Systems Research*, 3(3):429–457.

Vygotskij, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, Cambridge, Mass.

Společnost pro kognitivní vědu: komunita jako podmínka rozvoje kognitivní vědy v českém prostředí

Frederik Ďalak

Matematicko-fyzikální fakulta Univerzity Karlovy Matematický ústav UK
Email: frederik.dalak@matfyz.cuni.cz

Abstrakt

Společnost pro kognitivní vědu představuje studentskou a popularizační platformu, která v českém prostředí propojuje zájemce o mysl, poznávání a příbuzné obory. V situaci, kdy se kognitivní věda v České republice rozvíjí převážně napříč různými disciplínami a institucemi, mají podobné iniciativy zvláštní význam. Nejde pouze o organizaci přednášek nebo popularizačních aktivit, ale o vytváření komunity, která pomáhá udržovat pohromadě obor, jenž je svou povahou interdisciplinární a institucionálně rozptýlený. Cílem tohoto příspěvku je představit Společnost pro kognitivní vědu jako platformu, která propojuje studenty, odborníky i širší veřejnost a tím přispívá k větší viditelnosti a soudržnosti kognitivní vědy v českém prostředí.

1 Úvod

Společnost pro kognitivní vědu byla založena v roce 2011 na Vysoké škole ekonomické v Praze pod názvem Společ nost pro kognitivní vědu a filosofii. Za jejím vznikem stáli Ondřej Bečev a Petr Špecián, kteří usilovali o vytvoření prostoru pro setkávání lidí se zájmem o mysl, vědomí, jazyk, poznávání a současný výzkum v kognitivní vědě. Spolek od počátku pořádal přednášky a diskuse a po stupně rozšiřoval své aktivity i mimo Prahu, například do Brna. Zapojoval se také do širšího akademického a popularizačního dění, mimo jiné do organizace konference Kognice a umělý život a několika ročníků festivalu Na hlavu. Po roce 2018 však činnost spolku na několik let ustala. K jejímu obnovení došlo až v roce 2022, kdy se o nový rozběh zasadila nová generace studentů v čele s Karinou Zamrazilovou. Obnovený spolek navázal na původní myš lenku mezioborového setkávání, ale zároveň ji rozvinul směrem k současným formám studentské, popularizační a komunitní činnosti.
.

2 Fungování spolku

Spolek má v současnosti 21 členů a jeho těžištěm je pořádání pravidelných přednášek, popularizačních aktivit a neformálních setkání. Cílovou skupinou jsou především vysokoškolští studenti, ale také středoškoláci a širší ve řejnost. Vedle přednáškové činnosti se spolek věnuje i popularizaci nových poznatků na webu a sociálních sítích, kde přibližuje aktuální výzkum srozumitelnou formou. Důležitou součástí fungování jsou také pravidelná ne formální setkání nazývaná Cogtaily, jejichž cílem je propojovat aktivní i bývalé členy spolku a podporovat mezioborovou i mezigenerační diskusi. Spolek tak ne stojí jen na jednotlivých akcích, ale i na komunitním rozměru, který umožňuje dlouhodobé budování vztahů a spolupráce. Právě tento aspekt je pro mezioborový obor, jakým je kognitivní věda, mimořádně důležitý.



Obr. 1: Hlavní oblasti činnosti Společnosti pro kognitivní vědu.

3 Šíře témat a podoba oboru

Program spolku dobře ukazuje, jak široce lze dnes kogni tivní vědu chápat. V posledním období zazněly přednášky například o percepční teorii schizofrenie, o vztahu kognice, paměti a interpretace světa, o kognitivních mapách při prostorové orientaci osob se zrakovým postižením, o vztahu umělé inteligence a nevědomí nebo o pozornosti a vizuální percepci. Taková tematická různorodost ukazuje, že kognitivní věda dnes nepředstavuje jednu úzce vymezenou disciplínu, ale spíše prostor, v němž se setkávají psychologie, filosofie, neurovědy, biologie, jazykověda, informatika i technické obory. Právě tato rozmanitost je jednou z největších před ností kognitivní vědy, ale zároveň znamená, že podobný obor potřebuje místa, kde se jeho

různorodé části mohou skutečně propojovat. Spolek v tomto směru nevystupuje jen jako pořadatel akcí, ale i jako prostředí, které těmto různým tématům dává společný rámec.

4 Popularizace a veřejná viditelnost

Vedle přednášek je důležitou součástí činnosti spolku také popularizace současného výzkumu. Na webu společnosti se objevují texty věnované například psychedelikům a vědomí, neuroestetice, cocktail party efektu, vztahu AI a kreativity nebo dalším tématům na pomezí psychologie, neurověd a filosofie. Spolek zároveň využívá Instagram, YouTube a další online platformy k informování o ak cích, představování členů a sdílení záznamů vybraných přednášek. Tato činnost má význam nejen popularizační, ale i obo rový. Kognitivní věda je totiž v českém prostředí rozptý lená mezi více institucí a tradičních disciplín. Pokud má být vnímána jako smysluplný celek, je důležité vytvářet místa, kde se tento celek stává viditelným. Popularizace tak nepředstavuje jen doplněk odborné činnosti, ale i sou část toho, jak se obor představuje navenek a jak si buduje své místo v širším akademickém i veřejném prostoru.

5 Komunita jako podmínka rozvoje oboru

Jedním z důležitých aspektů činnosti spolku je vytváření komunity kolem kognitivní vědy. U mezioborových oblastí totiž často nestačí, aby vedle sebe pouze existovala různá témata a metody. Stejně důležité je, aby existoval prostor, ve kterém se mohou lidé z různých oborů pravidelně setkávat, navazovat spolupráci a společně přemýšlet o problémech, které žádná jednotlivá disciplína neobsáhne sama. Právě v tom lze vidět širší význam spolku. Nejde jen o studentskou organizaci v běžném smyslu slova, ale o plat formu, která pomáhá udržovat pohromadě obor, jenž je svou povahou interdisciplinární a institucionálně rozptý lený. Studie věnované interdisciplinárním oborům navíc ukazují, že pro jejich dlouhodobý rozvoj nejsou důležitá jen témata a metody, ale i infrastruktura, identita a ko munita, které těmto oborům dávají stabilnější podobu. Společnost pro kognitivní vědu tak nepomáhá pouze popu larizovat jednotlivá témata, ale zároveň podporuje vznik sdíleného prostoru, ve kterém lze kognitivní vědu vnímat jako skutečný celek.

6 Směřování do budoucna

Významnou součástí činnosti spolku zůstává také vazba na konferenci Kognice a umělý život, jejíž další ročník pro běhne v roce 2026. Právě propojení studentské iniciativy, popularizace a konferenčního prostředí ukazuje, že česká kognitivní věda nestojí jen na jednotlivých pracovištích, ale i na aktivní komunitě, která vytváří kontakty, sdílí témata a otevírá prostor pro nové spolupráce. Do budoucna se proto jako důležité jeví nejen pokračování v přednáškové a popularizační činnosti, ale také další roz šiřování členské základny, posilování veřejné viditelnosti oboru a propojování studentů s odborníky z různých in stitucí. Pokud má mít kognitivní věda v českém prostředí silnější a dlouhodobě udržitelnou podobu, nebude to zá viset pouze na jednotlivých výzkumných projektech, ale také na tom, zda kolem ní bude existovat aktivní komu nita, která ji bude dál rozvíjet a zpřítomňovat navenek.

7 Závěr

Společnost pro kognitivní vědu dnes nepředstavuje jen studentský spolek, ale platformu, která propojuje lidi, témata a instituce spojené s českou kognitivní vědou. Skrze přednášky, popularizační činnost, online komuni kaci i neformální setkávání vytváří prostor, ve kterém se mohou různé obory skutečně setkávat a vzájemně obo hacovat. Její význam proto nespočívá pouze v organizaci jednotlivých aktivit, ale především v budování komunity, která pomáhá dávat kognitivní vědě v českém prostředí konkrétní, viditelnou a dlouhodobě udržitelnou podobu.

Reference

Burns, T. W., O'Connor, D. J. a Stocklmayer, S. M. (2003). Science communication: A contemporary definition. *Public Understanding of Science*, 12(2):183–202.

Ku, S. T.-H. a Zehr, S. (2022). Disciplining interdisciplinarity: Infrastructure, identity, and interdisciplinary practice in nanoelsi research. *Science and Public Policy*, 49(5):765–780.

Pyrko, I., Dörfler, V. a Eden, C. (2017). Thinking together: What makes communities of practice work? *Human Relations*, 70(4):389–409.

Thagard, P. (2025). Cognitive science. Stanford Encyclopedia of Philosophy.

Would Psychedelics Alter the Fitness Landscape of Cooperation? An Experimental Framework for an Agent-Based Model

Aly Debevec-Kruse, Martin Takáč

Faculty of Mathematics, Physics and Informatics
Comenius University Bratislava
kruse4@uniba.sk, martin.takac@fmph.uniba.sk

Abstract

Personality psychology broadly recognises personality traits as strong predictors of behaviour. Empirical research has also demonstrated a clear relationship between personality traits and strategy selection in cooperative game play. Although personality traits remain relatively stable throughout an individual's lifetime, a limited number of interventions, such as cognitive behavioural therapy and mindfulness-based practices, have been shown to produce stable changes. However, more recently, the "psychedelic renaissance" has produced a growing pool of evidence linking psychedelic use to enduring shifts in personality. As regulatory frameworks surrounding psychedelics evolve, their use may become more widespread, raising questions about how large-scale personality changes could affect cooperative social dynamics. This paper aims to present a computational approach to studying cooperation. We review the literature on the effects of psychedelics on personality and the relationship between personality traits and cooperation, and based on these findings, we describe the conceptualisation of an agent-based simulation of multi-agent Prisoner's Dilemma interactions between agents with varying personality profiles. The proposed concept is intended as a starting point for future computational experiments aiming to investigate how varying levels of psychedelic use within a population may influence cooperative dynamics.

1 Introduction

Psychedelics are psychoactive substances that are well known for their unique pharmacological properties and capacity to alter the conscious experience (Kelmendi et al., 2022). Due to a rapid expansion of empirical research on their therapeutic potential, the last few decades have seen the emergence of a "psychedelic renaissance" (Hadar et al., 2023). Although promising, this renewed interest has placed increasing pressure on regulatory systems around the world to review restrictive legal frameworks. As a result, the global legal landscape surrounding psychedelics is undergoing a significant transformation, with various countries already moving toward decriminalisation and medicalisation (Monte et al., 2024). With these substances becoming more accessible, it becomes relevant to ask what their broader social impacts may be.

Within the area of Anthropology, the idea that psychedelics played an inherent role in the evolution of human cognition is beginning to take shape. Emerging empirical evidence that psychedelics can induce measurable and sometimes lasting changes in personality traits underlie this speculative concept (Rodríguez Arce and Winkelman, 2021). Given the established influence of personality traits on behaviour, it would be expected that reintroduction of psychedelics into a population on a large scale would produce individual behavioural changes, with downstream effects on social dynamics (Fleeson and Gallagher, 2009). However, the specific nature of these changes remains unclear and empirically answering these questions presents a challenge given the regulatory constraints surrounding these substances. Agent-based models (ABM) offer a promising solution, as their capacity for formal abstraction of complex systems allows for the systematic exploration of hypothetical scenarios. In this context, they can be used to examine the emergent patterns of social cooperation that can arise under conditions of large-scale psychedelic exposure.

Despite growing interest in both the cognitive impacts of psychedelic use and the role of personality in cooperative behaviour, to the best of our knowledge, there is currently no integrative model that captures the interaction between these domains. Existing models only incorporate personality in cooperation and are either overly complex or insufficiently aligned with the underlying mechanisms of interest. Although empirical studies provide valuable insights into the relationship between personality traits and cooperation, this literature remains fragmented and does not offer a clear, unified framework that can be directly operationalised in a modelling context. Therefore, the aim of this paper is to bridge this gap by reviewing the literature on (1) the effects of psychedelics on personality traits and (2) the relationship between personality and cooperative behaviour. Based on the reviewed literature, the paper seeks to develop a conceptual framework for personality-driven cooperative decision-making, designed for use in agent-based modelling. By doing so,

it aims to provide an initial step towards understanding how individual psychological changes may scale to influence broader social dynamics in contexts of increasing psychedelic use.

2 Related Work

Personality traits are understood as enduring patterns of thoughts, emotions, and behaviours that reflect consistent tendencies to respond in particular ways in different environments (Shum et al., 2025). A substantial body of research demonstrates that these traits can be measured with a high degree of reliability and stability throughout life and that they possess strong predictive validity in a wide range of life outcomes, including affect, life satisfaction, relationship stability, career success, and even longevity (Ozer and Benet-Martínez, 2006). More importantly, these personality measures, typically derived from questionnaire-based assessments, have been shown to correspond meaningfully with real-world behaviour (Fleeson and Gallagher, 2009). Within this domain, the Big Five personality traits (McCrae and John, 1992) have emerged as one of the most widely used and empirically supported frameworks for personality assessment. Its simplicity, broad applicability and cross-cultural generalisability make it particularly valuable for modelling purposes, as it captures the core structure underlying both lay and expert descriptions of personality. The model comprises five key dimensions, which together provide a robust and quantifiable representation of individual differences.

Trait	**Core Characteristics**
Extraversion	Sociability and positive emotionality
Agreeableness	Compassion and interpersonal trust
Conscientiousness	Productivity and goal-directed behaviour
Neuroticism	Emotional instability and susceptibility to negative affect
Openness	Curiosity, creativity, and receptiveness to novel experience

Tab. 1: The Big Five personality traits and their core characteristics adapted from McCrae and John (1992).

2.1 Psychedelics and Personality Traits

Inquiry into the role of psychedelics within personality psychology as a potential mechanism of behavioural change has become more prevalent. Empirical work has linked psychedelic use to enhanced prosocial behaviour, with personality traits representing one key domain. Several prospective studies have examined changes in personality after long-term use of psychedelics. The most consistent findings indicate a decrease in Neuroticism and an increase in Openness, with almost half of the reviewed studies supporting these changes in traits (Weiss et al., 2021, 2023; Godfrey et al., 2025). On the other hand, evidence of an increase in Extraversion and Agreeableness, while present, appears less consistent and replicable (Erritzoe et al., 2018; Weiss et al., 2023), and Conscientiousness shows little systematic change overall (Barrett et al., 2020; Weiss et al., 2023). Although these effects are often described as long-term, there is no consistent definition across studies; therefore, follow-up periods vary, ranging from weeks or months to multiple years, and may contribute to inconsistencies in findings. MacLean et al. (2011), for example, reported sustained changes in trait Openness one year after administration, while Godfrey et al. (2025) more recently supported the observed patterns of decreased Neuroticism and increased Agreeableness, but only measured personality changes one month post- administration. In addition to these considerations, the variation in compounds, including psilocybin, LSD, ayahuasca, and clinical and natural settings, may help explain the differences in observed outcomes. Psychedelic use can therefore be deemed most reliably associated with reduced Neuroticism and increased Openness, with more variable effects across other traits and study contexts.

In addition to this exploration of psychedelic-induced personality change, personality traits have also been found to influence both the likelihood of psychedelic use and the subjective quality of the experience itself. For example, greater Openness has been associated with more positive and meaningful experiences, and greater Neuroticism is associated with a more challenging trip (Kajonius et al., 2025). Similarly, patterns of use differ between personality profiles, with psychedelic users tending to be more open and extraverted and less neurotic than non-users, while interest in use among non-users is associated with greater Openness (Weiss et al., 2023). These more recent findings point to a more complex and dynamic interaction in which personality both shapes and is shaped by psychedelic use.

2.2 Personality Traits and Cooperation

As a substantial body of literature demonstrates the role of personality in shaping cooperative behaviour, cooperation represents a particularly important dynamic to investigate. Broadly defined as individuals or groups working together to achieve shared goals or mutual benefit (Parks et al., 2013), cooperation is often studied using controlled experimental paradigms employed in the form of social dilemma games (Mantas et al., 2022). The Prisoner's Dilemma is one of such social games, capturing the tension between individual self-interest and collective results in a minimal setting. It requires two individuals to independently decide whether

to cooperate or defect, with defection yielding the highest individual payoff when the other player cooperates (see Table 2). When iterated, decisions are no longer driven solely by immediate payoffs but also by expectations of future behaviour and accumulated outcomes, allowing cooperation to emerge over time. In this setting, strategies represent recurring patterns of behaviour throughout rounds, with Tit-for-Tat being one of the most prominent, initiating with cooperation and then mirroring the opponent's previous move (Mantas et al., 2022).

Tab. 2: Prisoner's Dilemma: Standard Payoff Matrix. (Holodny, 2016)

	B silent (coop.)	**B betrays A** (def.)
A silent (coop.)	Both: 1 year	A: 3 years, B free
A betrays B (def.)	A free, B: 3 years	Both: 2 years

Although these models are grounded in assumptions of rational decision-making, individuals often behave irrationally in social dilemmas (Fischbacher et al., 2000). These systematic deviations from predicted behaviour can be partially explained by individual differences such as personality traits (Ruch et al., 2017). Within cooperative game settings, this translates into observable differences in strategy selection, with personality traits contributing to variation in cooperative dispositions.

Among the Big Five, Agreeableness emerges as the most consistent predictor of cooperation, with higher levels associated with an increased likelihood of cooperative and prosocial behaviour across a range of experimental settings (Rustichini et al., 2019; Volk et al., 2011; Kagel and McGee, 2014). This has been associated with a preference for fairness and a reluctance to exploit others for personal gain (Zhao and Smillie, 2015). Openness has also been positively associated with cooperation as well as reciprocity, and may reflect greater strategic adaptability, where individuals are more willing to explore long-term cooperative strategies rather than prioritising immediate payoffs (Al-Ubaydli et al., 2016; Lönnqvist et al., 2011). In contrast, Conscientiousness is generally associated with lower levels of cooperation, which could reflect a more cautious, rule-oriented approach, in which individuals closely follow the structure of the game and favour defection as a means of avoiding exploitation (Rustichini et al., 2019; Al-Ubaydli et al., 2016; Lönnqvist et al., 2011). Extraversion shows a limited and context-dependent relationship with cooperation, with some evidence linking it to increased cooperative behaviour, particularly under incentivised conditions (Hirsh and Peterson, 2009). This may be explained by the rewarding nature of social interaction, although this interpretation remains less widely supported. Neuroticism has been associated with lower baseline cooperation and greater behavioural variability, which is believed to reflect increased sensitivity to risk and uncertainty, leading individuals to cooperate defensively to avoid retaliation or defect due to fear of exploitation (Al-Ubaydli et al., 2016; Lönnqvist et al., 2011; Hirsh and Peterson, 2009).

2.3 Agent-Based Models of Personality in Cooperation

Although existing work has explored the integration of personality traits into ABMs of cooperation, approaches remain varied in how personality is operationalised. Indraprastha (2011) incorporates personality as continuous traits, using Extraversion and Agreeableness to modulate attraction, repulsion, and group formation. Szilagyi (2003) defines personality in terms of behavioural update rules within multi-agent Prisoner's Dilemma settings. In this model, agents differ in how they adjust to cooperation by responding to past rewards (accountant), conforming to majority behaviour (conformist), or imitating the most successful neighbour (greedy). The most closely related and recent work, Li et al. (2026), uses the Big Five framework within a spatial Prisoner's Dilemma, integrating both personality traits and emotional dynamics. The model is comparatively complex, with a primary focus on modelling dynamic emotional states, with personality traits subsequently integrated, largely to determine strategy adaptation and selection.

3 The Agent-Based Model

Although the model proposed by Li et al. (2026) includes several relevant elements, its complexity makes it unsuitable as a baseline, as it limits the ability to isolate individual effects. As a primary aim of this project is to examine how changes in personality traits influence cooperation, a simpler and more modular baseline is preferable. Therefore, we propose an ABM that isolates the role of personality traits in shaping cooperative behaviour. The design follows a bottom-up approach, beginning with a simple probabilistic agent and incrementally introducing complexity. Personality is operationalised using the Big Five (OCEAN), with each trait influencing the probability of cooperation through weighted parameters informed by empirical literature. For example, Agreeableness, being consistently associated with increased cooperation in iterated Prisoner's Dilemma settings, would be assigned a positive weight, whereas Conscientiousness, linked to more cautious and less cooperative behaviour, would be assigned a negative weight. This structure would allow individual mechanisms to be introduced and tested independently,

ensuring that their effects remain separable and interpretable.

As behaviour is not stable across different social dilemma games, but is relatively stable across variations of the same game, the model focuses solely on the Prisoner's Dilemma (Haesevoets et al., 2018). This choice is further motivated by the availability of an extensive empirical literature, from which behavioural mechanisms can be extracted, as discussed above. The model would also adopt an iterated rather than one-shot structure, as repeated interactions allow behavioural patterns such as reciprocity and reputation to emerge (Pothos et al., 2011). The empirical relationships used to parametrise the model would be primarily derived from incentivised experimental settings. This is relevant, as personality effects can differ between hypothetical and incentivised tasks: for example, Agreeableness has been shown to predict lower cooperation in hypothetical settings but higher cooperation in incentivised ones (Hilbig et al., 2014). As ABMs are inherently payoff-driven, the use of incentivised findings provides a more appropriate empirical basis for parameterisation.

The model outputs would extend beyond simple cooperation rates and would additionally capture how different personality profiles influence cooperation over time, allowing both individual tendencies and population dynamics to be examined. Finally, to simulate the effects of psychedelic-induced personality change, the model would include a manipulation phase in which personality traits are adjusted for a proportion of the population. By varying the magnitude and distribution of these changes, the model would allow systematic exploration of how shifts in personality structure may influence cooperative dynamics.

4 Conclusion

We reviewed the literature on the effects of psychedelics on personality and the relationship between personality traits and cooperation, and proposed a simple framework for modelling personality-driven behaviour in an agent-based Prisoner's Dilemma. The aim was to isolate the role of personality traits using a modular and interpretable design grounded in empirical findings.

The proposed model provides a baseline that can be extended with additional complexity, such as a spatial structure, evolutionary dynamics or additional personality-related variables. It also allows for the simulation of psychedelic-induced personality change by shifting trait distributions within a population. Based on the literature, such changes would be expected to increase traits associated with prosocial behaviour, which may lead to higher and more stable levels of cooperation.

At the same time, behaviour is shaped by both individual traits and environmental factors, meaning that the model remains a reduction. Nonetheless, it provides a structured way to explore how changes in personality at the population level could influence cooperative dynamics.

Acknowledgement

This work was supported by the Slovak Society for Cognitive Science.

References

Al-Ubaydli, O., Jones, G. and Weel, J. (2016). Average player traits as predictors of cooperation in a repeated Prisoner's Dilemma. *Journal of Behavioral and Experimental Economics*, 64:50–60.

Barrett, F. S., Doss, M. K., Sepeda, N. D., Pekar, J. J. and Griffiths, R. R. (2020). Emotions and brain function are altered up to one month after a single high dose of psilocybin. *Scientific Reports*, 10:2214.

Erritzoe, D., Roseman, L., Nour, M. M., MacLean, K., Kaelen, M., Nutt, D. J. and Carhart-Harris, R. L. (2018). Effects of psilocybin therapy on personality structure. *Acta Psychiatrica Scandinavica*, 138(5):368–378.

Fischbacher, U., Gächter, S. and Fehr, E. (2000). Are people conditionally cooperative? Evidence from a public goods experiment. *SSRN Electronic Journal*.

Fleeson, W. and Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, 97(6):1097–1114.

Godfrey, K., Weiss, B., Zhang, X., Spriggs, M., Peill, J., Lyons, T., Carhart-Harris, R. and Erritzoe, D. (2025). An investigation of acute physiological and psychological moderators of psychedelic-induced personality change among healthy volunteers. *Neuroscience Applied*, 4:104092.

Hadar, A., David, J., Shalit, N., Roseman, L., Gross, R., Sessa, B. and Lev-Ran, S. (2023). The psychedelic renaissance in clinical research: A bibliometric analysis of three decades of human studies with psychedelics. *Journal of Psychoactive Drugs*, 55(1):1–10.

Haesevoets, T., Folmer, C. R., Bostyn, D. H. and Hiel, A. V. (2018). Behavioural consistency within the Prisoner's Dilemma game: The role of personality and situation. *European Journal of Personality*, 32(4):405–426.

Hilbig, B., Glöckner, A. and Zettler, I. (2014). Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology*, 107:529–539.

Hirsh, J. and Peterson, J. (2009). Extraversion, neuroticism, and the Prisoner's Dilemma. *Personality and Individual Differences*, 46:254–256.

Holodny, E. (2016). A professor explains how 1980s game theory could explain how morality evolved. https://www.sciencealert.com/a-professor-explains-how-1980s-game-theory-could-explain-how-morality-evolved Accessed: 2026-04-29.

Indraprastha, A. (2011). Computational model of social interaction in multiagent simulation based on personality traits. In *2011 International Conference on Advanced Computer Science and Information Systems*, pp. 183–188.

Kagel, J. and McGee, P. (2014). Personality and cooperation in finitely repeated Prisoner's Dilemma games. *Economics Letters*, 124(2):274–277.

Kajonius, P. J., Sjöström, D. and Claesdotter-Knutsson, E. (2025). Big Five personality and the psychedelic experience: An initial report. *Journal of Psychedelic Studies*, 9(4):412 – 416.

Kelmendi, B., Kaye, A. P., Pittenger, C. and Kwan, A. C. (2022). Psychedelics. *Current Biology*, 32(2):R63–R67.

Li, H., Li, C., Chen, M. et al. (2026). Cooperation promotion in the spatial Prisoner's Dilemma game through emotional and personality-based learning rules. *Complex Intelligent Systems*, 12:7.

Lönnqvist, J.-E., Verkasalo, M. and Walkowitz, G. (2011). It pays to pay – Big Five personality influences on co-operative behaviour in an incentivized and hypothetical Prisoner's Dilemma game. *Personality and Individual Differences*, 50(2):300–304.

MacLean, K. A., Johnson, M. W. and Griffiths, R. R. (2011). Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness. *Journal of Psychopharmacology*, 25(11):1453–1461.

Mantas, V., Pehlivanidis, A., Kotoula, V., Papanikolaou, K., Vassiliou, G., Papaiakovou, A. and Papageorgiou, C. (2022). Factors of influence in Prisoner's Dilemma task: A review of medical literature. *PeerJ*, 10:e12829.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Monte, A., Nicole, S., Joshua, B. et al. (2024). The rise of psychedelic drug use associated with legalization/decriminalization: An assessment with the nonmedical use of prescription drugs survey. *Annals of Emergency Medicine*, 83(3):283–285.

Ozer, D. J. and Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57:401–421.

Parks, C. D., Joireman, J. and Van Lange, P. A. M. (2013). Cooperation, trust, and antagonism: How public goods are promoted. *Perspectives on Psychological Science*, 14(3):483–497.

Pothos, E. M., Perry, G., Corr, P. J., Matthew, M. R. and Busemeyer, J. R. (2011). Understanding cooperation in the Prisoner's Dilemma game. *Personality and Individual Differences*, 51(3):210–215.

Rodríguez Arce, J. M. and Winkelman, M. J. (2021). Psychedelics, sociality, and human evolution. *Frontiers in Psychology*, 12:729425.

Ruch, W., Bruntsch, R. and Wagner, L. (2017). The role of character traits in economic games. *Personality and Individual Differences*, 108:186–190.

Rustichini, A., Proto, E. and Sofianos, A. (2019). Personality traits and economic preferences. *Journal of Political Economy*, 127(3):1459–1520.

Shum, K.-M., Ptaszynski, M. and Masui, F. (2025). Big Five personality trait prediction based on user comments. *Information*, 16(5).

Szilagyi, M. (2003). Simulation of multi-agent Prisoners' Dilemmas. *Systems Analysis Modelling Simulation*, 43:829–846.

Volk, S., Thöni, C. and Ruigrok, W. (2011). Personality, personal values and cooperation preferences in public goods games: A longitudinal study. *Personality and Individual Differences - PERS INDIV DIFFER*, 50.

Weiss, B., Nygart, V., Pommerencke, L. M., Carhart-Harris, R. L. and Erritzoe, D. (2021). Examining psychedelic-induced changes in social functioning and connectedness in a naturalistic online sample using the Five-Factor model of personality. *Frontiers in Psychology*, Volume 12 - 2021.

Weiss, B., Sleep, C. E., Beller, N. M., Erritzoe, D. and Campbell, W. K. (2023). Perceptions of psychedelic personality change, determinants of use, setting and drug moderation: Toward a holistic model. *Journal of Psychedelic Studies*, 7(3):200 – 226.

Zhao, K. and Smillie, L. (2015). The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review*, 19(3):277–302.

Associative Memory with Modern Hopfield Networks for Incremental and Open-Set Visual Learning

Jelena Epifanic, Kristína Malinovská

Faculty of Mathematics, Physics and Informatics
Comenius University Bratislava
epifanic1@uniba.sk, kristina.malinovska@fmph.uniba.sk

Abstract

In dynamic environments, robust representations and memory mechanisms are important for continuous learning and adaptation. In this study, we explore Modern Hopfield Networks (MHNs) as associative memory for incremental visual recognition. We propose a memory-based framework built on a pretrained self-supervised DINOv3 encoder with prototype-based or exemplar-based memory. Memory is updated incrementally and MHN-based refinement can optionally be applied to query embeddings before retrieval. For classification, we compare cosine similarity with MHN-based attention scoring. The method is evaluated under the OWOD split protocol using ground-truth object regions in order to focus on incremental and open-world recognition independently of object detection. The observed trade-off between known and unknown object recognition shows that the memory dynamics play a critical role in balancing stability and plasticity, where stronger associative retrieval improves familiarity-based classification but reduces sensitivity to novel or out-of-distribution inputs.

1 Introduction

Continuous learning in artificial intelligence is essential for real-world applications, where systems must autonomously acquire new knowledge, adapt to dynamic environments, and preserve previously learned information (Parisi et al., 2019). Biological systems serve as an inspiration, since humans are capable of learning incrementally and retaining stable representations over time despite changing inputs. In practice, ANNs still struggle with catastrophic forgetting, where new information interferes with earlier representations (Wang et al., 2023). In vision systems, some biological mechanisms such as stable representations and memory consolidation have been used as motivation for incremental learning.

To address this challenge, we explore Modern Hopfield Networks (MHNs) (Ramsauer et al., 2021) as an associative memory mechanism for open-world perception, combined with large-scale self-supervised visual representations. Our framework builds on a pretrained DINOv3 (Siméoni et al., 2025) encoder for feature extraction and supports both prototype-based and exemplar-based memory representations. This allows different strategies for storing and updating knowledge over time, depending on how much detail is preserved.

We investigate incremental visual recognition under the OWOD (Joseph et al., 2021) protocol. The model must continuously learn new object categories while still recognizing both known and previously unseen ones. We assume ground-truth object regions are available, so the focus is more on classification and open-set recognition rather than detection.

Inspired by biological learning principles, we try to balance plasticity and stability through explicit memory updates and retrieval. The results indicate that memory structure and update strategy may significantly affect retrieval quality, classification accuracy, and unknown class detection, though the effect varies across settings.

Our contributions are as follows:

- We present a memory-based framework for incremental and open-world visual recognition using MHN retrieval and frozen DINOv3 representations.
- We show that prototype-based and exemplar-based memory lead to different known–unknown trade-offs under incremental updates.
- We show that MHN-based retrieval improves known-class recognition, while stronger associative retrieval tends to reduce sensitivity to unknown categories.

2 Related work

Ramsauer et al. (2021) connect energy-based memory models with modern deep learning architectures. They revisit classical Hopfield networks by introducing continuous states. In this formulation, memory is represented as attractor states in an energy landscape.

The authors also show that MHNs are closely related to the softmax attention mechanism used in transformer architectures. From this perspective, attention

can be interpreted as a form of associative memory, where a query interacts with stored key–value patterns. This makes MHNs a natural bridge between energy-based models and large-scale representation learning.

In continual learning, these properties are particularly relevant, as the main challenge is balancing stability of previously acquired knowledge with plasticity for new information. Memory-based approaches are commonly used to reduce catastrophic forgetting by explicitly retaining or aggregating past experiences (Kudithipudi et al., 2022). From this perspective, MHNs provide a natural mechanism for associative memory retrieval, which can be interpreted as a computational abstraction of similarity-based recall processes in cognitive systems.

Similar ideas appear in prototype- and exemplar-based representations. Prototype models compress experience into a central representation per class, while exemplar-based models preserve individual instances and rely on similarity-based retrieval over stored memories. Different granularity representations originate from cognitive science as competing models of human categorization. Prototypes capture central tendencies, while exemplars preserve instance-level variability (Nosofsky, 1992). Rather than being fundamentally different mechanisms, more recent work suggests they can be viewed as different regimes of a shared similarity-based retrieval process (Smith, 2014).

This interpretation naturally aligns with Modern Hopfield Networks. MHNs operate over an energy landscape where stored patterns act as attractors. Depending on retrieval sharpness, the same mechanism can behave more like a compressed prototype system or a distributed exemplar system. MHN-based retrieval naturally spans both regimes by operating directly on similarity in representational space.

Despite these connections, the role of MHNs as an associative memory mechanism in incremental open-world visual recognition remains underexplored. In particular, it is not well understood how memory granularity affects stability–plasticity trade-offs in continual learning settings. To investigate this behavior, we evaluate our method under the OWOD protocol (Joseph et al., 2021). We follow the standard split setup on the COCO dataset (Lin et al., 2014), where object-level annotations are incrementally introduced as new classes. This provides a test-bed for continual learning with expanding label spaces and varying memory representations.

Existing OWOD methods such as OW-DETR (Gupta et al., 2022) focus on object detection and unknown object discovery. In contrast, we assume ground-truth object regions and focus on the recognition problem. Our approach also differs from standard parametric learning methods, since knowledge is stored and updated in memory while the DINOv3 encoder remains frozen.

3 Our method

Motivated by biological continual learning systems that maintain both stable and adaptable representations through memory consolidation and incremental updates, we propose a memory-based framework for incremental visual recognition. The proposed approach combines large-scale self-supervised visual representations with an associative memory mechanism. We use a pretrained DINOv3 encoder for feature extraction and a memory module parameterized as either class prototypes or class exemplars. The pipeline consists of: (i) feature extraction from object regions, (ii) supervised memory construction from labeled support data, (iii) incremental memory updates that allow new classes to be integrated without retraining from scratch, and (iv) open-world inference with optional MHN-based query refinement and unknown rejection.

3.1 Memory Construction

Given an input image and corresponding object regions (from annotations), each region is encoded using a pretrained self-supervised DINOv3 backbone (Siméoni et al., 2025). The encoder outputs a feature embedding $z \in \mathbf{R}^d$, which is L2-normalized before being stored in memory:

$$\hat{z} = \frac{z}{\|z\|_2}. \tag{1}$$

Let $M = \{(p_j, c_j)\}_{j=1}^{N_p}$ denote memory support items, where p_j is normalized prototype or exemplar embedding and c_j its associated class label. We consider two memory granularities:

Prototype-based memory. For each class, a fixed number of memory items is initialized (e.g., k-means centroids or selected examples).

Exemplar-based memory. Support embeddings are stored directly without aggregation preserving instance-level detail, though at higher memory cost.

This design reflects a trade-off between compact semantic representation and fine-grained instance preservation.

Memory is built from labeled support samples grouped by class. For each class, support embeddings are added to memory according to the selected representation strategy. Class-wise confidence statistics are maintained for open-set thresholding:

$$\tau_c = \mu_c - k\sigma_c, \tag{2}$$

where μ_c and σ_c are class c confidence mean and standard deviation, and k is a tunable margin parameter.

These thresholds are later used for unknown rejection during inference.

3.2 Incremental Memory Update

To support continual learning, memory is updated incrementally as new labeled samples arrive. In prototype-based memory, updates follow a stability-preserving exponential moving average (EMA):

$$p \leftarrow \text{norm}\left((1-\alpha)p + \alpha\hat{z}\right), \tag{3}$$

applied when the cosine similarity between the incoming sample and the nearest prototype exceeds a predefined update threshold ($\tau_{\text{update}} = 0.7$). If the similarity is below a second threshold ($\tau_{\text{new}} = 0.5$), a new prototype is appended to memory. Samples with intermediate similarity are ignored. In exemplar-based memory, new embeddings are directly stored without modification. This improves continuous adaptation while preserving past class representations.

3.3 Inference

At inference time, query embeddings can optionally be refined using a Modern Hopfield retrieval step over memory items (Ramsauer et al., 2021).

Given a query embedding $\hat{z}$ and memory matrix P, retrieval is computed as:

$$w = \text{softmax}(\beta\hat{z}P^\top), \quad r = wP, \tag{4}$$

where β controls the retrieval sharpness.
The query is then refined via interpolation:

$$\tilde{z} = \text{norm}\big((1-\lambda)\hat{z} + \lambda r\big), \tag{5}$$

where λ controls the influence of retrieved memory. When MHN-based query refinement is enabled, the refined embedding $\tilde{z}$ is used for classification. Otherwise, the original embedding $\hat{z}$ is used. This step is optional and introduces an associative recall mechanism before classification, where two different strategies are employed:

Cosine-based classification. The predicted class is assigned using nearest-neighbor search in cosine similarity space. Since the focus of this work is on memory representations and associative retrieval, cosine similarity is used as a simple baseline to isolate the contribution of MHN-based memory mechanisms. Similar nearest-neighbor evaluation is used for DINO-style representations, but is typically reported on image classification benchmarks (e.g., ImageNet) in a closed-set setting rather than under open-world recognition protocols.

MHN-based classification. When both refinement and MHN classification are enabled, Eq. (4) is applied first to $\hat{z}$ and then again to $\tilde{z}$.

Class scores are obtained by summing attention weights over memory items belonging to each class. The predicted class corresponds to the class with the highest aggregated score. The confidence is defined as the normalized attention mass of the predicted class, i.e., the fraction of total attention assigned to that class. In this setting, the parameter m acts as a global minimum-confidence threshold.

Following the open-world setting described earlier, a prediction is accepted as known only if its confidence score exceeds a class-dependent threshold (in MHN mode, optionally a global minimum-confidence term m). Otherwise, the sample is assigned to an unknown category.

4 Experiments

4.1 Dataset and Protocol

We follow the standard task-wise split setting (Gupta et al., 2022) on the COCO dataset (Lin et al., 2014), with four incremental stages ($t \in \{1, 2, 3, 4\}$), and report results on the corresponding test partitions. We use ground-truth object bounding boxes as input regions and do not evaluate object proposal or detection performance. All methods are evaluated at the region level, where each ground-truth object crop is treated as an independent recognition instance.

Unknown recognition uses the class-aware confidence threshold as defined in Eq. 2. In MHN mode, an additional minimum-confidence constraint m is applied. We further evaluate the effect of memory representation (prototype vs exemplar), classifier type (cosine vs MHN), and MHN query refinement.

Performance is evaluated on ground-truth instances under the OWOD incremental protocol. We report known-class accuracy, unknown recall (U-Recall), and open-set error rate (OSE).

4.2 Experimental Setup

We evaluate configurations varying along three factors: (i) memory representation (prototype vs. exemplar), (ii) classification strategy (cosine similarity as a baseline vs. MHN-based classification), and (iii) optional MHN-based query refinement.

This results in a total of eight configurations, summarized in Table 1. All setups are evaluated under the OWOD protocol across tasks T1–T4.

All experiments use frozen DINOv3 (Siméoni et al., 2025) features extracted from object regions. Memory is built from labeled support samples (n_{support}) using either: (i) prototype-based aggregation, or (ii) direct exemplar storage.

During training, optional refinement updates memory using additional labeled samples. This allows gradual adaptation without full retraining.

Unless stated otherwise, we keep the encoder, data splits, and extracted features fixed across all experiments, and vary only memory and inference configurations. Key hyperparameters include the number of

Row	Memory	Classifier	Refinement
1	Prototype	Cosine (baseline)	–
2	Prototype	MHN	–
3	Prototype	Cosine	MHN
4	Prototype	MHN	MHN
5	Exemplar	Cosine (baseline)	–
6	Exemplar	MHN	–
7	Exemplar	Cosine	MHN
8	Exemplar	MHN	MHN

Tab. 1: Evaluated configurations across memory representation, classifier type, and optional MHN-based query refinement.

support samples ($n_{\text{support}} = 10$), number of prototypes, update strategy (`ema` or `append only`), and the unknown-threshold parameter k. To ensure stability of the exemplar memory, we limit storage to at most 100 samples per class.

5 Results

For the first experiment, we keep the following parameters fixed: the MHN sharpness parameter $\beta = 30.0$, the unknown threshold parameter $k = 0.5$, and the minimum confidence threshold $m = 0.35$.

For exemplar-based memory (Table 2), MHN-based classification improves known-class accuracy compared to the cosine baseline (from 7.84% to 19.65%). A further improvement is obtained when MHN-based query refinement is included (up to 25.33%). However, this gain is accompanied by a strong decrease in unknown-class performance, where U-Recall drops from 81.53% to 33.74% and OSE increases accordingly. A similar behavior is observed for prototype-based memory. MHN classification improves known-class accuracy (61.98% to 66.74%) and slightly increases unknown recall. However, MHN-based refinement has a strong negative effect on open-set behavior, reducing U-Recall to 3.19% in the cosine+refine setting and significantly increasing OSE. The results are consistent with the idea that the prototype-based regime enforces stronger abstraction, while the exemplar-based regime preserves variability necessary for open-set discrimination.

The low accuracy of the exemplar cosine baseline (Table 2) is likely caused by the limited number of stored exemplars per class. Due to the high variability of object instances in COCO, the memory may not sufficiently cover the feature space of each category, making nearest-neighbor retrieval less robust than prototype-based representations. This interpretation is consistent with the results in Table 5, where increasing the number of stored exemplars improves performance.

Overall, these results suggest that MHN-based retrieval improves recognition performance on known classes, but also introduces a bias toward predicting known categories (Fig. 1). This effect becomes more pronounced when MHN query refinement is applied. Interestingly, despite this bias, the person class still shows a tendency to be misclassified as unknown, indicating class-specific ambiguity.

Method	Known Acc.	U-Recall	OSE
Exemplar-based memory			
Cosine NN (baseline)	7.84	81.53	18.47
MHN classification	19.65	72.97	27.03
Cosine + MHN refine	19.97	51.29	48.71
MHN + refine	25.33	33.74	66.26
Prototype-based memory			
Cosine (baseline)	61.98	25.30	74.70
MHN classification	66.74	38.54	61.46
Cosine + MHN refine	66.27	3.19	96.81
MHN + refine	65.75	29.45	70.55

Tab. 2: Comparison of exemplar- and prototype-based memory under cosine and MHN-based inference variants.

When we compare the incremental tasks, the same pattern remains stable, as shown in Table 3. Earlier tasks achieve higher known-class accuracy, while later tasks show a gradual shift toward a more balanced but still unstable known–unknown trade-off.

Task	Known Acc. (%)	U-Recall (%)	OSE (%)
T1	82.52	23.80	76.20
T2	70.03	35.64	64.36
T3	59.08	56.17	43.83
T4	55.34	N/A	N/A

Tab. 3: Per-task performance for prototype-based memory with MHN classification across OWOD incremental stages T1–T4.

At Task 4, all 80 COCO categories in the OWOD protocol have already been introduced. Since no unseen categories remain in the evaluation set, unknown-class metrics such as U-Recall and OSE cannot be computed. For this reason, only known-class accuracy is reported for Task 4.

We also designed the experiments to analyze the effect of individual parameters on model behavior. In this setting, we increase the MHN sharpness parameter to $\beta = 60.0$, the unknown threshold to $k = 0.7$, and the minimum confidence threshold to $m = 0.6$.

The results, summarized in Table 4, show how these changes affect performance across tasks T1–T4. Increasing β makes the model more confident but significantly reduces unknown recall, leading to strong over-assignment to known classes. Increasing k makes the decision boundary slightly more permissive for un-

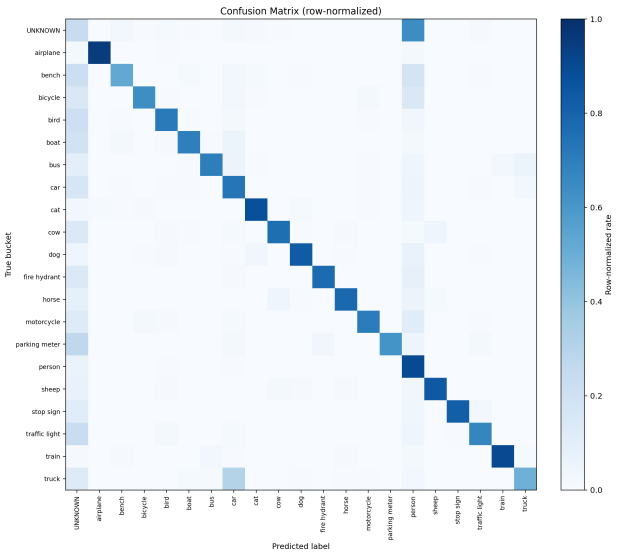


Fig. 1: Confusion matrix of MHN-based classification. The model shows strong diagonal structure for known classes, but a tendency to assign unknown samples to known categories.

known detection. Increasing the minimum confidence threshold m improves open-set behavior by rejecting uncertain predictions, but at the cost of lower known-class accuracy. This pattern is consistent with the attractor-based memory systems, where increased attractor strength can lead to overgeneralization toward familiar categories at the expense of sensitivity to novel inputs (Hopfield, 1982).

Task	Known Acc. (%)	U-Recall (%)	OSE (%)
Base	66.74	38.54	61.46
$k = 0.7$	67.03	36.95	63.05
$\beta = 60$	68.67	8.83	91.17
$m = 0.6$	52.13	65.30	34.70

Tab. 4: Prototype-based memory with MHN classification under different parameter settings. The baseline (Base) refers to the default configuration used in the main experiments, with $\beta = 30$, $k = 0.5$, and $m = 0.35$.

The MHN sharpness parameter β was selected through preliminary hyperparameter exploration over a range of candidate values, from nearly uniform retrieval ($\beta \approx 0$) to highly selective retrieval ($\beta = 60$). Increasing β increases the sharpness of associative retrieval. In our experiments, $\beta = 60$ produced highly selective memory retrieval and substantially reduced unknown recall, while $\beta = 30$ provided a more balanced trade-off between known- and unknown-class recognition. Higher values concentrate attention on the most similar memory items, whereas lower values average information over a larger set of prototypes. For the settings considered here, $\beta = 30$ provided a more balanced trade-off than $\beta = 60$.

The results in Table 5, further highlight the role of memory granularity in the stability–plasticity trade-off. Increasing the number of stored exemplars per class leads to a substantial improvement in unknown recall (from 33.74% to 52.51%), while only slightly reducing known-class accuracy. This indicates that a denser exemplar memory improves the model's ability to reject unseen categories, likely by providing a more complete coverage of the known-class feature space. Since exemplar memory construction heavily depends on the available support examples, this aspect requires further investigation.

Task	Known Acc. (%)	U-Recall (%)	OSE (%)
MHN + refine (baseline)	25.33	33.74	66.26
50 examples/class	22.76	39.40	60.60
100 examples/class	22.38	52.51	47.49

Tab. 5: Exemplar-based memory with MHN classification. Mean performance across tasks for different exemplar settings, with 20 examples per class as the baseline.

To better understand the MHN mechanism, a qualitative analysis was conducted. The DINO mask visualization shows a clean part-level structure on both the dog and the person, distinguishing it from the background. This confirms that DINOv3 features encode meaningful semantics, which aligns with levels of human organization in figure–ground segmentation. Interestingly, the cosine and MHN attention maps highlight different regions (head/torso/limbs). MHN attention focuses more on the head and central parts of the body, showing stronger weighting on salient features.

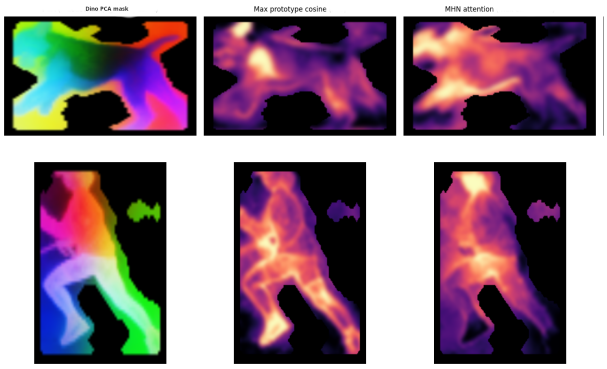


Fig. 2: Dense DINO Feature and Memory Retrieval Visualization

6 Conclusion

In this work, we studied how MHN-based associative memory supports incremental open-world visual recognition. Different memory representations originating from cognitive theories, classification strategies, and optional MHN query refinement are studied under the OWOD protocol. The presented results are obtained using ground-truth object regions. In a full OWOD system, object proposals would introduce additional noise and localization errors. Evaluating the proposed memory framework together with object detection remains an important direction for future work. Using the OWOD protocol on COCO also allows future integration with existing detection-based benchmarks. Unlike standard parametric approaches, the proposed framework updates knowledge through memory mechanisms while keeping the visual encoder fixed. This provides an alternative perspective on incremental learning based on retrieval and memory dynamics rather than continual parameter optimization.

The results show a clear trade-off between known-class recognition and open-set performance. While similarity-based retrieval can improve known-class accuracy, it also tends to increase the bias toward known categories, especially when query refinement is used. This leads to lower unknown-class recall and higher open-set error. Stronger associative retrieval increases attraction toward familiar memory states, while reducing sensitivity to novel inputs. We also find that memory structure has an important effect. Exemplar-based memory improves unknown detection when more samples per class are stored, suggesting that better coverage of the feature space helps rejecting unseen categories. In contrast, more compact memory representations are more stable for known classes but less effective for open-set recognition.

Overall, the results show that retrieval-based memory is a simple and effective approach for open-world recognition, but its behavior depends strongly on memory design and retrieval settings.

Acknowledgement

The presented work is based on the research conducted for a master's thesis entitled *Incremental and Open-Set Visual Learning with Cognitively Inspired Memory Mechanisms* under the supervision of Dr. Kristína Malinovská (coauthor) and Prof. Markus Vincze (ACIN, ETIT, TUW, Vienna).

This paper was written at the Centre for Cognitive Science at DAI FMPI Comenius University Bratislava, with the support of the Slovak Society for Cognitive SSKV[1].

[1] https://cogsci.fmph.uniba.sk/sskv/

References

Gupta, A., Narayan, S., Joseph, K. J., Khan, S., Khan, F. S. and Shah, M. (2022). Ow-detr: Open-world detection transformer. *arXiv*.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Joseph, K. J., Khan, S., Khan, F. S. and Balasubramanian, V. N. (2021). Towards open world object detection. *arXiv*.

Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J., Brna, A. P., Raja, S. C., Cheney, N., Clune, J., Daram, A., Fusi, S., Helfer, P., Kay, L., Ketz, N., Kira, Z., Kolouri, S., Krichmar, J. L., Kriegman, S. and Levin, M. (2022). Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P. (2014). Microsoft coco: Common objects in context. *arXiv*.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In *Essays in Honor of William K. Estes*, vol. 1, pp. 149–167.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M. and Hochreiter, S. (2021). Hopfield networks is all you need. *arXiv*.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A. and Tolan, J. (2025). Dinov3. *arXiv*.

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin Review*, 21:312–331.

Wang, L., Zhang, X., Su, H. and Zhu, J. (2023). A comprehensive survey of continual learning: Theory, method and application. *arXiv*.

Unsupervised Topographic Representations of Motor Primitives in Humanoid Manipulation and Gestures

Radovan Gregor, Igor Farkaš

Department of Applied Informatics, Comenius University Bratislava
{radovan.gregor, igor.farkas}@fmph.uniba.sk

Abstract

Understanding the computational basis of action recognition is a key challenge in social cognition as well as human–robot interaction. Inspired by the mirror neuron system, we propose a two-level architecture for motor primitive discovery and processing, applied to the humanoid robot. In the first level, two self-organising maps learn topologically structured manifolds over arm kinematics and hand kinematics covering seven manipulation and gesture actions. The second level uses the Echo State Network that encodes the temporal grammar of primitive sequences to predict actions and infer intentions from partial observations. In this paper, we focus on the first level and report preliminary results.

1 Introduction and Related Work

The ability to recognise observed actions and infer the intentions behind them is fundamental to social interaction. An observer who can interpret partner's motor behaviour can proactively adapt, anticipate, and collaborate more effectively. The Mirror Neuron System (MNS) offers a compelling biological model for this capacity: mirror neurons discharge both when an individual executes a motor act and when it observes another performing the same act (Gallese et al., 1996), grounding action understanding in an internal motor vocabulary rather than abstract perceptual classification (Rizzolatti and Craighero, 2004). Neurophysiological studies have shown that parietal mirror neurons encode not only the observed action but also its goal (Fogassi et al., 2005). This goal-specificity is a key feature of the action-understanding hypothesis and motivates the design of computational systems that can similarly distinguish actions at the level of motor primitives and at the level of goal-directed sequences (Oztop et al., 2006).

The classical computational approach encodes observed motion directly from kinematic or visual signals using recurrent architectures trained end-to-end on action labels (Shahroudy et al., 2016). More recent methods replace recurrence with transformer-based architectures, enabling the modeling of long-range temporal dependencies through self-attention (Bertasius et al., 2021). However, neither approach embeds the biological insight that action understanding operates across two levels simultaneously: a repertoire of elementary motion primitives and the goal-directed sequences they compose (Wolpert et al., 2003). Computational models of the MNS have addressed this by modelling the mirror system as a self-organising map (SOM) (Kohonen, 1982) that receives both motion and context inputs, showing that goal-specific neuron pools emerge naturally from the geometric relationships between these inputs without any explicit goal representation (Thill et al., 2011). The same principle underpins chain models of intention understanding (Chersi et al., 2011; Erlhagen et al., 2006), in which sequences of motor primitives activate a learned chain that predicts the overall action goal. In our previous MNS-related work, we also exploited topographic mapping as an organizing principle (Pospíchal et al., 2019; Rebrová et al., 2013).

Here we focus on topographic representation of proprioceptive signals whose sequence describes the moving arm. Proprioception is one of the least understood senses in cognitive neuroscience, yet fundamental for the control of movement. There is evidence that topography of cortical representations is a widespread organising principle of the brain (deCharms et al., 2000). It has also been proposed for proporioception at the level of latent representations of the variational autoencoder (Grogan et al., 2024).

We propose a two-level MNS-inspired architecture grounded in these findings, using topographic organisation at the lower level. There, two SOMs discover a vocabulary of reusable motion primitives from arm and hand kinematics independently – a dual-stream design consistent with the observation that motion and context inputs to parietal mirror neurons originate from distinct neural pathways (Thill et al., 2011). The topology-preserving projection naturally groups temporally similar movement episodes into prototype nodes, which serve as a discrete primitive vocabulary (Hemeren and Thill, 2011). At the higher level, an Echo State Network (Lukoševičius and Jaeger, 2009) encodes the temporal ordering of these primitives into state-level representations from which actions and intentions can be inferred, extending our previous work on arm trajectory prediction (Gregor et al., 2025).

Our central research question is: How much of a robot's motor sequence structure is determined by motion primitives alone, and how much additional contex-

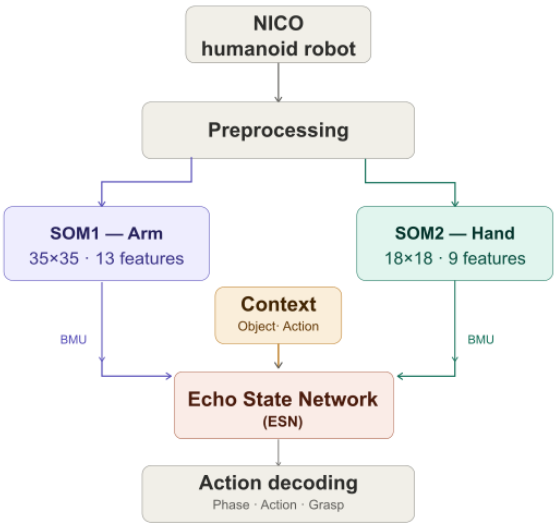


Fig. 1: The overall concept of the presented work.

tual information (object properties, action intent) is required for accurate motion prediction?

2 Data and Methods

2.1 Simulation and Data Collection

To build the dataset for motion primitive discovery, we employ a physics-based Unity simulation environment featuring NICO robot (Kerzel et al., 2017), a humanoid platform for multi-modal interaction. In our setup, the robot performs 7 manipulation and gesture actions: $Pick, Eat, Place, Push, Tap, Point, Wave$. Each action is performed on a library of 14 physically diverse objects spanning spheres, cubes, cylinders, discs, and irregular shapes in multiple sizes. Object manipulation actions $Pick, Eat, Place$ are executed with either a Power or Precision grasp, while the remaining actions each employ a dedicated hand configuration. To ensure the diversity in the dataset, the object positions are randomized within a reachable peri-personal workspace, requiring the robot to generate a varying motion trajectories. Motion trajectories are generated through inverse kinematics, while 13 finger joints are controlled via four synergy parameters inspired by human grasping studies. Movement execution is organized as phase-based state machines containing 27 movement phases across the complete motion cycle. Representative phase labels include $preshape, movetograsp, closing, lift.$ Motion is recorded at 60 Hz. The resulting dataset contains 4,200 simulated trials, of which 4,106 completed successfully, yielding approximately 1.28 million frames. Following preprocessing and window extraction, 61,523 balanced windows were used for SOM training.

2.2 Feature Analysis and Selection

Prior to SOM training, a feature selection pipeline was applied to 108 candidate signals to identify and remove redundancy and features that are workspace-dependent and do not generalise across object locations. Based on correlation analysis and hierarchical clustering we selected 22 non-redundant features, which are divided into two functional sets: (1) 13 arm features for arm SOM (A-SOM): 3 TCP (Tool Center Point) palm velocities, 6 arm joint velocities, palm orientation quaternion; and (2) 9 hand features for hand SOM (H-SOM): aperture and aperture velocity, 3 thumb and middle finger positions, 4 finger joint velocities.

2.3 Windowing and Normalisation

Motion logs were segmented into overlapping windows of 30 frames (0.5 s at 60 Hz) with a stride of 15 frames (50% overlap). Each window was represented by its per-feature mean value. Phase and action labels were assigned from the central frame of the window. To prevent over-represented actions from dominating SOM organization, windows were stratified such that each action contributed an equal number of samples. The final balanced dataset contained 8,789 windows per action. Features were independently normalized using z-score computed from training-set statistics.

2.4 SOM Training

Two independent SOMs (Kohonen, 1982) were trained using a custom implementation with a Gaussian neighbourhood function, initial learning rate $\alpha_0 = 0.5$ decaying exponentially to $\alpha_f = 0.01$, and initial neighbourhood radius $\sigma_0 = \max(R, C)/2$ decaying to $\sigma_f = 1.0$ over $T = 100$ epochs. Each SOM maintains a weight matrix $\mathbf{W} \in \mathbb{R}^{R \times C \times d_s}$, where $R \times C$ is the grid size and d_s is the SOM-specific feature dimension ($d_A = 13$ for A-SOM, $d_H = 9$ for H-SOM). The winner (best matching unit, BMU) for input $\mathbf{x}_t$ is the neuron with minimal Euclidean distance to the weight vector. A-SOM uses a 35×35 grid, while H-SOM uses an 18×18 grid. Grid sizes were selected empirically. Increasing A-SOM from 25×25 to 35×35 improved phase purity from approx. 68% to 74% while maintaining high neuron utilization. For H-SOM, larger grids provided little benefit because the synergy controller produces only a small number of discrete hand configurations. Training was repeated with three random seeds to assess sensitivity to initialisation. A-SOM phase purity varied by ± 0.08 percentage points (pp) across runs and H-SOM grasp purity by ± 0.31 pp, confirming that both maps converge to stable representations independent of initialisation.

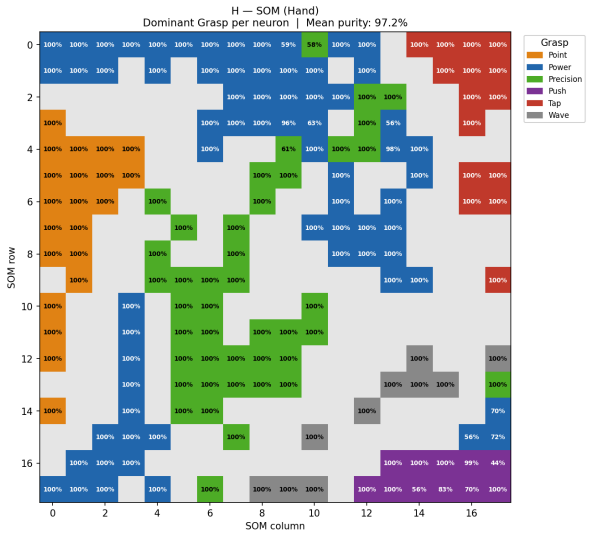


Fig. 2: Topographic organisation (in H-SOM) of dominant grasp types across neurons.

2.5 Evaluation Metrics

To quantify map organization, we measured the purity of phase, action and grasp. For each active neuron, the dominant class label was identified. Purity was then computed as the proportion of samples belonging to that dominant label, averaged across active neurons. Phase purity measures the consistency of movement-phase representation, action purity measures action specificity, and grasp purity measures grasp-type specificity. Neuron utilization was computed as the percentage of neurons that became BMUs for at least one training sample. Dead neurons are defined as neurons that never become the BMU for any training sample. In H-SOM, dead neurons primarily occur between isolated hand-configuration clusters and therefore serve as separators between distinct configuration regions rather than indicating poor training.

3 Results

A-SOM learned to use up to 90% of its neurons (as winners for any input). Phase purity (74%) is the strongest label dimension, with large coherent zones emerging without supervision. The map encodes a directional reaching topology: approach phases occupy the left half and post-action phases the right, a gradient learned entirely from kinematics with no temporal supervision. Action purity (67%) and grasp purity (66%) are lower, as expected, since arm kinematics alone cannot distinguish actions sharing identical reach trajectories, nor encode grip type, both of which are delegated to H-SOM.

H-SOM achieves 51.5% neuron utilisation, lower than A-SOM by design – the hand synergy controller produces discrete, non-overlapping configuration spaces, so dead neurons act as hard boundaries between isolated action islands rather than representing wasted capacity (see Fig. 2). Grasp purity (97.2%) is the best result overall. Point, Tap, Wave, and Push each form perfectly pure isolated clusters, whereas Power and Precision diverge into distinct, non-overlapping regions, with impurities restricted to a narrow transition band in the mid-close range. Each hand pose occupies its own island with no kinematic path between them, reflecting the hard discreteness of the synergy controller. Action purity (72.9%) and phase purity (76.1%) confirm that finger-level signals carry complementary discriminative structure to the arm-level primitives of A-SOM.

4 Planned Upper-level Architecture

The lower-level SOM outputs serve as the input representation for an Echo State Network (ESN). Trajectories will be encoded as temporal sequences of BMUs, providing a discrete motor-primitive vocabulary learned from arm and hand kinematics. Extending our prior work on trajectory prediction shown in Figure 3 (Gregor et al., 2025) the upper level will employ an ESN-based forward model trained on motion primitive sequences. The system will learn to predict future motion states from short observation windows, leveraging the unsupervised SOM representations of arm and hand kinematics. By encoding sequences of SOM activations, the ESN is expected to capture the temporal grammar underlying goal-directed actions and provide a compact state representation for subsequent action recognition and intention inference.

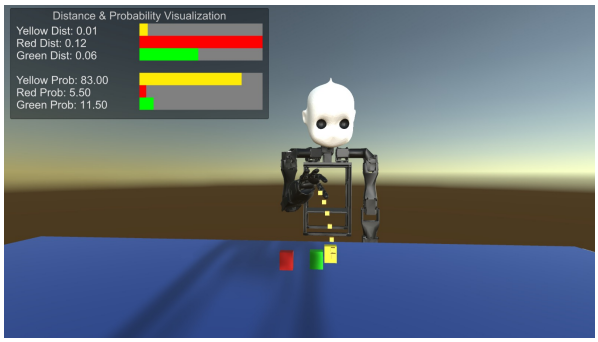


Fig. 3: ESN prediction of future arm trajectories.

5 Conclusion

We have presented a two-level MNS-inspired architecture for motion primitive discovery in a humanoid robot, with focus on the first level. Two SOMs trained on arm kinematics and hand kinematics jointly encode a primitive vocabulary with clear topographic structure. At the same time, the results suggest that arm and hand

kinematics encode complementary aspects of motor behaviour – "where and how" of reaching vs. "what and how" of grasping – not recoverable from either stream alone. This complementarity of SOMs may reflect the dual representational strategy attributed to the MNS, where elementary motor primitives and their sequential composition operate at distinct levels of abstraction.

A key limitation is that the SOM is currently trained on balanced, stratified windows from a single observer executing its own actions. Extension to cross-embodiment observation – where NICO observes a human or a different-size robot – will require the mirroring mechanism to resolve the viewpoint and embodiment gap. This is consistent with MNS-inspired computational accounts but requires direct empirical validation beyond the scope of the present work.

Acknowledgment

This research was supported by Slovak Research and Development Agency, project APVV-21-0105.

References

Bertasius, G., Wang, H., and Torresani, L. (2021). Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, pages 813–824.

Chersi, F., Ferrari, P. F., and Fogassi, L. (2011). Neuronal chains for actions in the parietal lobe: A computational model. *PLOS ONE*, 6(11).

deCharms, R. C., , and Zador, A. (2000). Neural representation and the cortical code. *Annual Review of Neuroscience*, 23:613–647.

Erlhagen, W., Mukovskiy, A., and Bicho, E. (2006). A dynamic model for action understanding and goal-directed imitation. *Brain Research*, 1083:174–188.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science*, 308:662–667.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119:593–609.

Gregor, R., Farkaš, I., Malinovska, K., Sobota, B., and Kerzel, M. (2025). Mirroring-based prediction of robot arm movements using an echo state network. In *International Conference on Automation, Control and Robots*, pages 49–54.

Grogan, M., Blum, K. P., Wu, Y., Harston, J. A., Miller, L. E., and Faisal, A. A. (2024). Predicting proprioceptive cortical anatomy and neural coding with topographic autoencoders. *PLOS Computational Biology*.

Hemeren, P. E. and Thill, S. (2011). Deriving motion primitives through action segmentation. *Frontiers in Psychology*, 1:243.

Kerzel, M. et al. (2017). NICO – Neuro-Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 113–120.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Oztop, E., Kawato, M., and Arbib, M. A. (2006). Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19:254–271.

Pospíchal, J., Farkaš, I., Pecháč, M., and Malinovská, K. (2019). Modeling self-organized emergence of perspective in/variant mirror neurons in a robotic system. In *IEEE International Conference on Development and Learning*.

Rebrová, K., Pecháč, M., and Farkaš, I. (2013). Towards a robotic model of the mirror neuron system. In *IEEE International Conference on Development and Learning and on Epigenetic Robotics*.

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019.

Thill, S., Svensson, H., and Ziemke, T. (2011). Modeling the development of goal-specificity in mirror neurons. *Cognitive Computation*, 3(4):525–538.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B*, 358(1431):593–602.

Multi-Scale Simulation of Synaptic Plasticity Induced by Repetitive Magnetic Stimulation

Nicholas Hananeia

Centrum kognitívnych vied, Comenius University Bratislava
Email: nicholas.hananeia@fmph.uniba.sk

This paper presents a concise summary of previously published work describing a multi-scale model of location- and frequency-dependent synaptic plasticity induced by repetitive magnetic stimulation (Hananeia et al., 2025).

Abstract

Repetitive transcranial magnetic stimulation (rTMS) induces long-term synaptic changes, but the mechanisms underlying these effects remain unclear. We developed a biophysically realistic modelling framework that integrates a voltage-dependent plasticity model with neuronal compartment models and electric field simulations to predict repetitive magnetic stimulation (rMS)-induced excitatory synaptic plasticity.

The framework reproduced experimentally observed LTP-like plasticity in hippocampal CA1 pyramidal neurons following 10 Hz repetitive magnetic stimulation, including strong distance dependence and localisation to proximal synapses. Simulations also predicted reduced plasticity at lower stimulation frequencies (5 Hz and 1 Hz). In addition, the model replicated distal synaptic plasticity induced by local electrical theta-burst stimulation and predicted both proximal and distal plasticity during magnetic theta-burst protocols. These responses were strongly facilitated by dendritic spikes and showed limited sensitivity to inhibitory suppression.

This framework enables high-resolution simulation of rMS-induced synaptic plasticity and may improve parameter screening and the design of more effective plasticity-inducing stimulation protocols.

1 Introduction

Transcranial magnetic stimulation has been used clinically for decades, yet its cellular mechanism remains elusive.

TMS induces strong electric fields in the brain tissue beneath the stimulation coil. These fields primarily excite axonal compartments rather than somatic or dendritic membranes, generating action potentials that propagate to the soma and dendrites. The clinical effect is thought to arise from its modulation of neuroplasticity; however, the relationship between the electric fields and long-term synaptic plasticity (long-term potentiation (LTP) and long-term depression (LTD)) remains incompletely understood.

We developed a modelling framework that allows simulation of realistic magnetic stimulation in morphologically complex cells with high spatiotemporal resolution. This allows modelling of plasticity arising from both global (direct electric field induced extracellular voltage) and local (events such as NMDA spikes and large postsynaptic potentials) effects.

Studies (Lenz et al., 2015) have found strong distance dependence in the response of synapses to magnetic stimulation, wherein proximal synapses are potentiated far more strongly than distal ones by a 10 Hz magnetic stimulation protocol.

Using a reduced morphology model of the CA1 pyramidal cell, we were able to reproduce the distance-dependent pattern of magnetic stimulation induced LTP. Additionally, we predict frequency-dependent potentiation, with higher frequencies producing stronger effects. Finally, we found that high-frequency theta-burst magnetic stimulation of 100 Hz can induce potentiation in distal tuft synapses.

2 Methods

All simulations were implemented in NEURON using the NeMo-TMS toolbox (Shirinpour et al., 2021) with a biophysically detailed but morphologically reduced CA1 pyramidal cell model (Tomko et al., 2021). This reduced model has many of the morphological features typical of a CA1 pyramidal cell, but is far less complex than a complete reconstruction. It has two basal dendrites that each have two branches, as well as a single apical dendrite with a thickened trunk that branches into two thin tufts. Additionally, it has three oblique branches connecting directly to the apical dendrite (Fig. 1).

Because magnetic stimulation initiates action potentials in axonal locations, an explicit axon was included in the model. This consisted of an initial segment, an unmyelinated terminal segment, and myelinated internodal segments. Myelinated internodes were included to enable reliable magnetic-stimulation-induced firing; rMS-initiated action potentials do not occur without myelination.

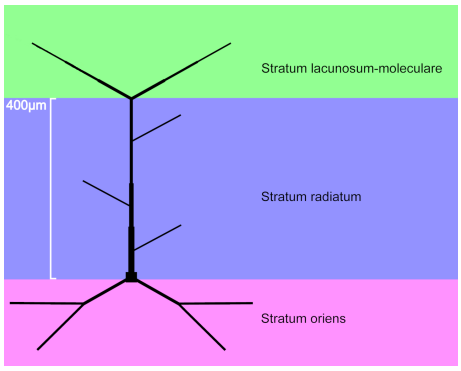


Fig. 1: Schematic of reduced model of CA1 pyramidal cell with layers highlighted.

The model has many ion channels, and is tuned to reproduce the electrophysiological behavior of the biological cell, and its behavior is independently validated by the HippoUnit test suite (Sáray et al., 2021).

128 excitatory AMPA/NMDA synapses and 18 inhibitory GABA synapses were placed in all layers of the cell and assigned to the various layers in the model (see Fig. 1), based on relative synapse numbers in Megıas et al. (2001).

To model synaptic plasticity, we used the synaptic plasticity model of Ebner et al. (Ebner et al., 2019) for AMPA/NMDA synapses and the non-plastic Exp2Syn method with a negative reversal potential was used for GABA synapses.

The Ebner model is a voltage-based phenomenological plasticity model of a combined AMPA/NMDA synapse with internal variables corresponding to biophysical processes. It considers four separate pathways of plasticity: presynaptic LTP, presynaptic LTD, postsynaptic LTP, and postsynaptic LTD, which govern two separate internal weights, presynaptic and postsynaptic.

The effective synaptic weight, which determines EPSP amplitude, is defined as the product of the presynaptic and postsynaptic weights. All reported weight changes refer to this effective synaptic weight.

Simulation of the rMS-induced magnetic field was done using the NeMo-TMS toolbox. The toolbox's uniform magnetic field option was chosen for simplicity as we modelled an in-vitro experiment, where complex geometry of neural tissue was not a factor. Here, the magnetic stimulation is implemented as an extracellular potential at every compartment in the model, shown in equation 1, where V_e is the extracellular voltage, E is the electric field vector, and ds is the differential over a given segment of membrane, and $E_{(x,y,z)}$ are the x, y, and z components of the electric field at location (x, y, z):

$$V_e(x,y,z) = -\int \vec{E}\cdot d\vec{s} = -\vec{E}\cdot\vec{s} = -(E_x x + E_y y + E_z z) \quad (1)$$

The temporal component is encoded in a separate timecourse which is multiplied by this uniform spatial component; the time course also contains the temporal shape (monophasic or biphasic) of the stimulus. These temporal shapes are based on recorded waveforms from real stimulators and are included in the NeMo-TMS framework.

When simulating an rMS pulse we stimulated all synapses simultaneously with the extracellular voltage application, with a delay of 1ms between the magnetic stimulus and the synaptic activation. This was chosen because rMS strongly activates axon terminals without the activation of their cell bodies, which would lead to synaptic activation.

3 Results

3.1 Validation of the synaptic plasticity model for local electrical simulation

Since our synaptic plasticity model was not tuned for the CA1 region, we used the experiments of Ikegaya et al. (Ikegaya et al., 2002). In this experiment, the cells were directly electrically stimulated by a series of 900 pulses delivered to the Schaffer collaterals (stratum radiatum and stratum oriens targets) at a constant frequency of either 1, 10, 30, or 100 Hz. In these experiments, weight changes were measured both with inhibition present and absent.

After adjusting the parameters of our model, we found results within observed standard deviation for all conditions except for the 1 Hz and 10 Hz control cases. The change in synaptic weights was negligible at frequencies below 10 Hz (Fig. 2).

3.2 Proximal LTP induction by repetitive magnetic stimulation

To determine whether our plasticity model would be applicable to a realistic rMS protocol, we simulated the experiments of Lenz et al. (Lenz et al., 2015). In this study, an rMS of 900 pulses at 10 Hz was applied to hippocampal slice cultures. We modeled this with a train of 900 monophasic pulses with an amplitude of 275 V/m, which is above the cell's firing threshold.

As a result, the mean Schaffer collateral weight increased by 30%, with the LTP induction strongly localized in the proximal dendrites, consistent with what was observed in Lenz. Considering only the proximal dendrites of the stratum radiatum, we observed an average weight increase of 65% above baseline, whereas the distal synapses on the apical dendrite showed no LTP.

To predict the frequency dependence of synaptic plasticity induced by this rMS protocol, we also delivered 900 stimuli at lower frequencies of 1, 5, 6, 8, and 9 Hz. Proximal LTP was induced, with the amount of potentiation much less at lower frequencies (Fig. 3).

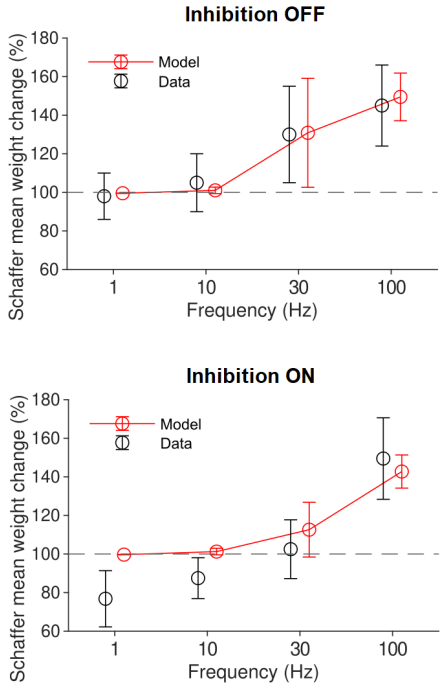


Fig. 2: Model reproduces electrical stimulation data from Ikegaya et al. for most frequencies. Top: Frequency dependence with inhibition off. Bottom: Frequency dependence with inhibition on.

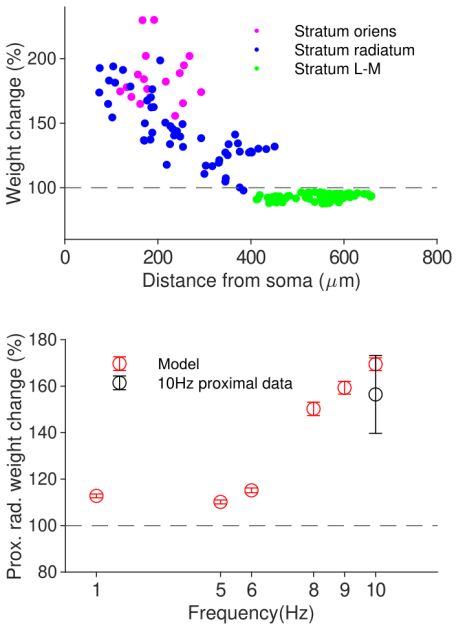


Fig. 3: Potentiation from 10 Hz magnetic stimulation is much stronger in the proximal dendrites than distal, and is stronger with higher frequencies. Top: Distance dependence of weight change in response to 10 Hz 900 pulse stimulus. Bottom: Frequency dependence of average weight change in proximal (<200 μm) synapses.

Consistent with data, our simulations showed that rMS induced LTP at much lower frequencies (10 Hz vs 30-100 for classical protocols) than commonly used plasticity inducing protocols.

3.3 Distal LTP induction by theta-burst stimulation protocols

To determine what conditions could show LTP elicitation at greater distances, we implemented a simulation of a theta-burst stimulation protocol.

Induction of LTP in these distal synapses has been shown to require the presence of dendritic spikes (Kim et al., 2015). Since our plasticity model is capable of generating responses to subthreshold local events, we investigated if our model could reproduce this.

We applied a local electrical stimulation protocol of 15 bursts of 5 pulses at 100 Hz, applied only to the synapses in the distal tuft. We repeated the same protocol both with and without inhibition present, since in the study of Kim et al. (Kim et al., 2015), inhibition was blocked to facilitate dendritic spikes.

As a comparison to these local electrical stimuli, we investigated the effect of rMS applied in a similar theta-burst fashion. In these cases, the rMS-like stimulus was a biphasic pulse delivered at 275 V/m. Because this was modelling magnetic stimulation, all synapses were activated simultaneously with the magnetic pulse, similarly to in Section 3.2.

We compared the LTP/LTD distance profiles of all four cases - local electric stimulation and magnetic stimulation both with and without inhibition. With purely electric TBS and no inhibition, LTP was observed in the distal tuft, along with the presence of dendritic spikes. With inhibition present, the dendritic spikes were abolished, and LTP greatly reduced (Fig. 4).

With the magnetic stimulation, however, we observed LTP not only in the proximal branches, but in the distal tuft, both in the presence and absence of inhibition. In these cases, proximal LTP was of similar magnitude to distal LTP, unlike in our 10 Hz simulations (Fig. 5).

4 Discussion and Conclusion

We present a modelling framework capable of reproducing experimentally observed location-dependent synaptic plasticity induced by repetitive magnetic stimulation. The model replicated the key finding of distance-dependent potentiation of proximal synapses, with decreasing synaptic change with increasing distance.

Furthermore, we predict a dependence of potentiation magnitude on stimulation frequency, with stronger synaptic potentiation at higher frequencies.

Finally, the model suggests that distal tuft synapses can be potentiated by theta-burst stimulation:

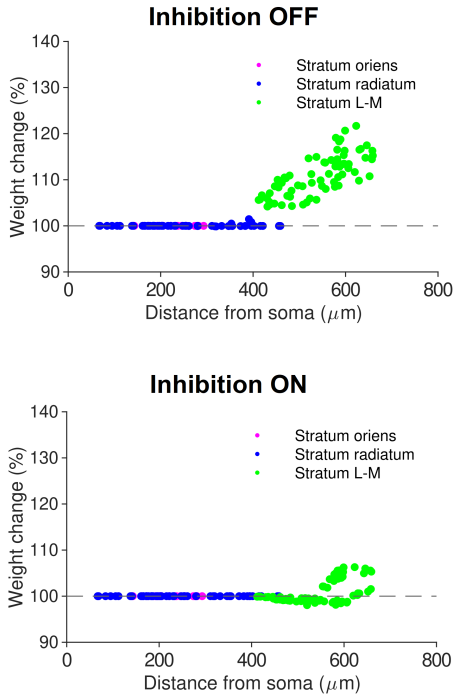


Fig. 4: Electrical theta-burst stimulation produces LTP in distal tuft when inhibition disabled. Top: Inhibition off. Bottom: Inhibition on.

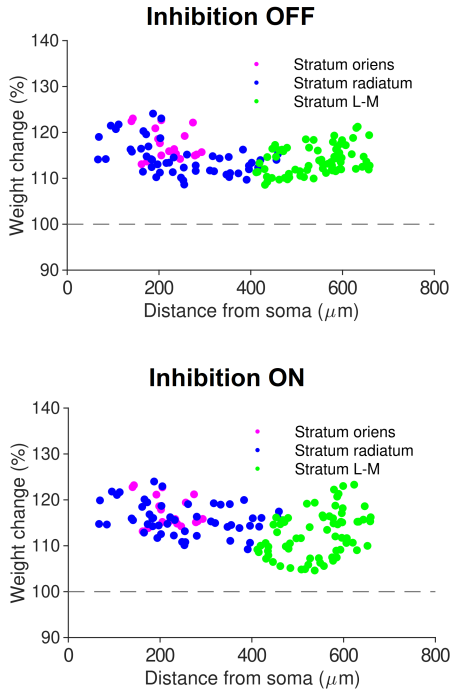


Fig. 5: Magnetic theta-burst stimulation produces LTP all over the cell both with and without inhibition. Top: Inhibition off. Bottom: Inhibition on.

locally, through dendritic spike induction, and magnetically, through a similar burst protocol.

Our voltage-dependent model explains these observations: simpler spike-dependent plasticity models would not capture attenuation of backpropagating action potentials. Inhibition can further modulate this process, as inhibitory postsynaptic potentials reduce the efficacy of backpropagating action potentials. Similarly, dendritic spikes mediated by voltage-gated sodium channels are often subthreshold but, in our model, were crucial for distal tuft potentiation under theta-burst stimulation.

Limitations of the model include the absence of metaplasticity (Abraham, 2008) and inhibitory plasticity (Lenz et al., 2016), both of which may influence responses during prolonged stimulation. In particular, disinhibition is a known mechanism for facilitating excitability and plasticity (Lenz a Vlachos, 2016). We also employed a reduced single-cell morphology rather than a complete cell reconstruction and did not model network-level effects, which are likely relevant given the large tissue volumes stimulated by TMS. These remain important directions for future work.

Overall, this study shows that biophysically grounded modelling can reproduce known effects of repetitive magnetic stimulation while generating testable predictions for new protocols. Such approaches may help guide future experimental design and optimisation of stimulation paradigms.

Acknowledgements

This work was completed at Justus Liebig University Giessen under the supervision of Prof. Peter Jedlička.

References

Abraham, W. C. (2008). Metaplasticity: tuning synapses and networks for plasticity. *Nature Reviews Neuroscience*, 9(5):387–387.

Ebner, C., Clopath, C., Jedlicka, P. a Cuntz., H. (2019). Unifying long-term plasticity rules for excitatory synapses by modeling dendrites of cortical pyramidal neurons. *Cell Reports*, (29.13):4295–4307.

Hananeia, N., Ebner, C., Galanis, C., Cuntz, H., Opitz, A., Vlachos, A. a Jedlicka, P. (2025). Multi-scale modelling of location-and frequency-dependent synaptic plasticity induced by repetitive magnetic stimulation in the dendrites of pyramidal neurons. *PLoS computational biology*, 21(11):e1012295.

Ikegaya, Y., Ishizaka, Y. a Matsuki., N. (2002). Bdnf attenuates hippocampal ltd via activation of phospholipase c: implications for a vertical shift in the frequency–response curve of synaptic plasticity. *European Journal of Neuroscience*, (16.1):145–148.

Kim, Y., Hsu, C.-L., Cembrowski, M. S., Mensh, B. D. a Spruston., N. (2015). Dendritic sodium spikes are required for long-term potentiation at distal synapses on hippocampal pyramidal neurons. *eLife*, (4):e06414.

Lenz, M., Galanis, C., Müller-Dahlhaus, F., Opitz, A., Wierenga, C. J., Szabó, G., Ziemann, U., Deller, T., Funke, K. a Vlachos, A. (2016). Repetitive magnetic stimulation induces plasticity of inhibitory synapses. *Nature Communications*, 7(10020).

Lenz, M., Platschek, S., Priesemann, V., Becker, D., Willems, L. M., Ziemann, U., Deller, T., Müller-Dahlhaus, F., Jedlicka, P. a Vlachos, A. (2015). Repetitive magnetic stimulation induces plasticity of excitatory postsynapses on proximal dendrites of cultured mouse ca1 pyramidal neurons. *Brain Structure and Function*, (220.6):3323–3337.

Lenz, M. a Vlachos, A. (2016). Releasing the cortical brake by non-invasive electromagnetic stimulation? rtms induces ltd of gabaergic neurotransmission. *Frontiers in Neural Circuits*, 10(96).

Megıas, M., Emri, Z. S., Freund, T. F. a Gulyas., A. I. (2001). Total number and distribution of inhibitory and excitatory synapses on hippocampal ca1 pyramidal cells. *Neuroscience*, (102.3):527–540.

Sáray, S., Rössert, C. A., Appukuttan, S., Migliore, R., Vitale, P., Lupascu, C. A., Bologna, L. L., Van Geit, W., Romani, A., Davison, A. P. et al. (2021). Hippounit: A software tool for the automated testing and systematic comparison of detailed models of hippocampal neurons based on electrophysiological data. *PLoS Computational Biology*, 17(1):e1008114.

Shirinpour, S., Hananeia, N., Rosado, J., Tran, H., Galanis, C., Vlachos, A., Jedlicka, P., Queisser, G. a Opitz, A. (2021). Multi-scale modeling toolbox for single neuron and subcellular activity under transcranial magnetic stimulation. *Brain Stimulation*, 14(6):1470–1482.

Tomko, M., Benuskova, L. a Jedlicka, P. (2021). A new reduced-morphology model for ca1 pyramidal cells and its validation and comparison with other models using hippounit. *Scientific Reports*, 11(7615).

Virtual reality enabling cognitive map acquisition for persons with visual impairment: A Systematic Literature Review

Kryštof Havlík and Miroslav Macík

Czech Technical University in Prague, Faculty of Electrical Engineering
Jugoslávských partyzánů 1580/3, Praha
Email: havlikr4@fel.cvut.cz

Abstract

A cognitive map (CM) is a mental model comprising information about an environment and is used to infer knowledge for spatial cognition tasks. Humans acquire a CM integrating information perceived by their senses. This makes the acquisition process different for persons with visual impairment (PVI) as they cannot easily perceive detailed information about an environment from a distance. CM can be acquired even from exploring a virtual representation of an environment. Virtual reality (VR) is a promising tool for research of CM acquisition, as it can facilitate the process, is highly customizable and allows to measure many features of the experience. This article presents a systematic literature review (SLR) of 21 recent papers focusing on VR systems which facilitate acquisition of CM for PVI. The review aims to answer four research questions aimed at mapping the current state of VR systems for PVI concerning the techniques used, the best practices identified and how are they evaluated. Suggestions for functional and non-functional requirements of new potential VR system are synthesized from the findings of the review.

1 Introduction

Like all humans, persons with visual impairment create a mental model of an environment called a cognitive map. CM contains knowledge such as locations, layout, routes, distance, and directions between locations, which can be used to infer information to support spatial tasks. Appropriate CM supports a PVI's ability to move independently, efficiently, and safely in an environment as shown by Chebat et al. (2020). CM can be acquired through direct experience of the environment, but also through indirect methods, which allow the acquisition before visiting the real environment, avoiding potential dangers associated with exploring an unknown environment. CMs acquired through a combination of indirect and direct experiences have also been observed to enable better execution of spatial tasks than CMs acquired solely through direct exploration, as indirect methods have been shown to be better suited for acquisition of allocentric (object to object) spatial knowledge by Ottink et al. (2022).

Virtual reality is one of the indirect acquisition methods, but it has not been fully explored for PVI. With increased affordability of VR technology such as head mounted displays in the recent years, more experiments aiming to create a VR system suitable for PVI are emerging. VR offers great conditions for research as it allows researchers to have complete control over an environment in which experiments are conducted, making it an attractive choice for facilitating both CM acquisition and evaluation.

2 Review goals

The motivation for this literature review is to map out the current state of research of cognitive map (CM) acquisition by visually impaired persons (PVI) and virtual reality (VR) for PVI. The results will serve as one of the sources based on which a VR framework that would support CM acquisition for PVI will be designed. This paper describes a subpart of the review which focuses specifically on recent VR systems which enable acquisition of spatial information for PVI.

The review aims to answer the following research questions, which were created using the CIMO scheme. Rationales behind the questions are presented under them:

RQ1: What techniques can be used to create a VR system for PVI?

rationale: Identify the available base building blocks for the VR framework.

RQ2: Which features of a VR system improve CM acquisition for PVI?

rationale: Identify the requirements on the VR framework so that it facilitates CM acquisition.

RQ3: To which extent are cognitive maps acquired in VR usable in the real environment?

rationale: Verify if the framework is useful for potential real life application.

RQ4: Which methods can be used to evaluate qualities of a CM for PVI?

rationale: Identify what methods to consider to use for evaluation and also if any of them could be integrated into the VR framework.

3 Review methodology

The review followed the PRISMA guideline, starting with defining the research topic and research questions presented in the previous part. Based on those a database search was conducted in the identification phase. The results from the search were further filtered in the screening phase and later judged for eligibility. The steps of the process can be seen in the workflow diagram in fig 1.

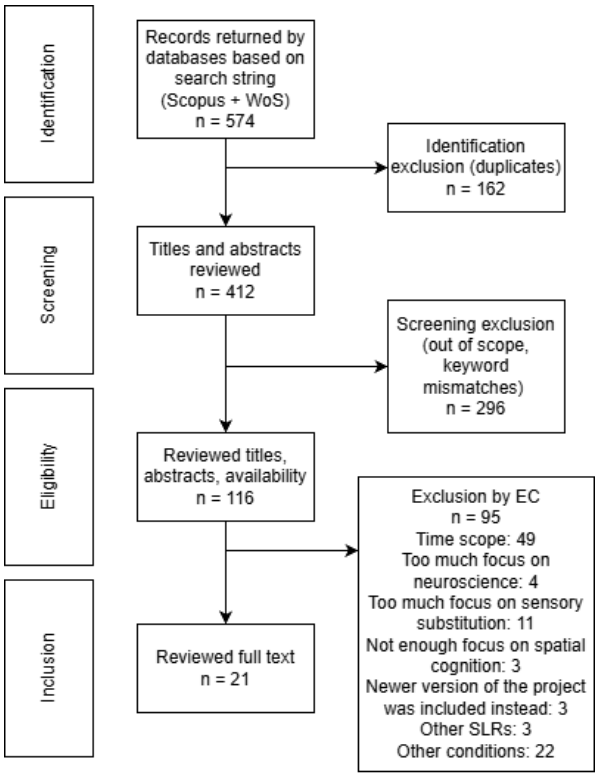


Fig. 1: SLR flow chart

3.1 Identification

The search string was created from the research topic *Supporting cognitive map acquisition among persons with vision impairments in VR*, which was divided into three blocks: 1. Cognitive map, 2. Persons with vision impairments and 3. Virtual reality. From these blocks a document-term matrix was constructed, which can be seen in table 1. This table was then used to construct the following search string:

("cognitive map" OR "mental map*" OR ("spatial" AND ("cognition" OR "knowledge" OR "learning")) OR "survey knowledge" OR "route knowledge" OR "indoor navigation" OR wayfinding OR pathfinding OR "orientation and mobility" OR "O&M") AND ("visual impairment*" OR "visual disability*" OR "visually impaired" OR blindness OR blind) AND ("virtual reality" OR VR OR "extended reality" OR XR OR "augmented reality" OR AR OR "mixed reality" OR MR)*

For the search, the following digital libraries Scopus, Web of science, SpringerLink, ACM and IEEE Xplore were considered. In the end only Scopus and Web of science were used, as the results from SpringerLink were focused too much on medicinal papers and the search string was too restrictive for ACM and IEEE Xplore. The search was conducted on the 1st of April 2026.

The identification phase ended with automated duplicates removal using the Zotero software for paper cataloging, with total of 162 duplicates being removed.

3.2 LLM assisted Screening

Since Title and Abstract Screening is one of the most time-consuming parts of an SLR, a decision was made to try to automate it using large language models (LLM). With the recent rise of LLMs, there have been some research into using them for various parts of SLRs, mainly for the screening and identification phases. In their paper Joos et al. (2026) reached very promising results in Title and Abstract Screening using a consensus of multiple LLMs. Joos et al. also freely provide the software they developed for the LLM screening called LLMSurver [1], which was used for screening.

For the screening models provided by e-INFRA CZ were used as they were free and without rate limits for academic purposes. The models chosen were qwen 3.5, deepseek-v3.2-thinking, gpt-oss-120b and redhatai scout, as they were the built for reasoning and long form tasks from among the other available models provided by e-INFRA CZ.

In the beginning a random sample of 20 papers was selected and manually screened. This manual screening was used as a goal result for iterative prompt and consensus scheme refinement. The resulting scheme was, that a paper was included when at least two models agreed on inclusion.

Validation on a manually screened sample of 40 papers yielded a precision of 0.90 and recall of 1.00, indicating that the approach effectively reduced manual screening effort while maintaining high sensitivity.

3.3 Filtering the subpart

The last exclusion of papers was done to filter out papers focusing on recent VR systems which enable acquisition of spatial information for PVI. In this phase papers older than 10 years and those which did not focus on VR (but only on sensory substitution devices) or spatial cognition were excluded. This part was done manually by reviewing titles, abstracts and full texts.

[1] https://llmsurver.dbvis.de/

Tab. 1: Document-term matrix

	Cognitive map	Persons with vision impairments	Virtual reality
Related terms	cognitive map acquisition, mental map, spatial cognition, spatial knowledge, spatial learning, survey knowledge, route knowledge	visual impairment, visually impaired, visual disability, blindness, blind	VR, virtual environment
Broader terms	indoor navigation, wayfinding, pathfinding, Orientation and Mobility, O&M	(without vision), nonvisual	Extended reality (XR), Augmented reality (AR), Mixed reality (MR)

3.4 Included papers

After the exclusion process 21 papers were left, whose full texts were reviewed. The papers are listed in tables 2 and 4.

4 Results and research questions discussion

The contents of the papers were analyzed with the goal of answering the research questions. The findings concerning each one are presented in the following parts.

4.1 RQ1: What techniques are used to create a VR system for PVI?

Analysis of the reviewed systems identified five design dimensions: movement, user input, output modality, output triggers, and additional accessibility features. Table 2 summarizes the implementations. The percentage representation can be found in table 3.

The systems reviewed in the SLR were split in the implementation of movement between stationary design and real walking. All the systems used audio for output and two thirds used it in combination with haptic features. The most used interaction method is avatar collision at 47.62% alluding to the use of contact with object for orientation by PVI.

Around a third of the systems used already available VR technology such as head mounted displays with controllers and treadmills modified for PVI, prominent example being Han et al. (2025). A quarter of the implementations aimed for simplicity and affordability using only a smartphone as both an input and output device and for position tracking. Several experiments tried to create their own specialized devices, such as the *Canetroller* by Siu et al. (2020) a white cane connected to a harness simulating both a texture of an object with vibrations and hard contact using brakes preventing movement of the cane. Some systems like *The virtual Cane* in Lahav (2022) try to create a more modified version of reality which incorporate sensory substitution methods to give the user detailed information about an object from a distance, which is normally unavailable to them.

4.2 RQ2: Which features of a VR system improve CM acquisition for PVI?

4.2.1 Movement

In the case of movement Han et al. (2025) observed better performance in obstacle detection with free walking, while using a VR treadmill improved layout learning in perpendicular environments. Kubota et al. (2025) showed that real body turning improved navigation of complex, non-perpendicular routes. No paper unfortunately compared real walking and stationary setups.

4.2.2 System output and modalities

In terms of output modality Siu et al. (2020) and Zhao et al. (2024) highlight the benefits of using a multimodal feedback in enabling CM acquisition at multiple levels of detail at the same time, increasing CM acquisition efficacy. Kubota et al. (2025) reported that audio feedback is better for general orientation while haptic feedback is better for conveying finer object details. Others like Kunz et al. (2018) and Guerrón et al. (2020) warn not to overload one modality with too much information, as the user will not be able to process all the information.

Andrade et al. (2021) demonstrated echolocation as an effective method for spatial layout acquisition. Ecological validity of sounds—particularly footsteps and voice—was important to users, with unnatural audio causing distraction Andrade et al. (2021); Ochiai et al. (2025). May et al. (2020) identified that key landmarks should be highlighted with more salient cues

4.2.3 Additional features

Cobo et al. (2018) observed that access to information about distant objects improved spatial confidence and localization . Lahav et al. (2018); Lahav (2022) found that distal exploration yielded more allocentric CM models and greater object detail recall than real environment exploration, though at the cost of prolonged exploration time. Sonification and voice cues for objects in the user's line of sight accelerated and improved exploration according to Cobo et al. (2017). Simpler, more intuitive interfaces were preferred by PVI in papers by Guerrón et al. (2020); Lahav et al. (2018).

Tab. 2: Types of VR techniques implemented across the reviewed.

Article	Movement	User input	Output	Output trigger	Additional features
Andrade et al. (2021)	stationary, fixed, controller	head tracking, controller	spatial audio, physics sim., echoes	footsteps loop, on demand	teleport
Cobo et al. (2017)	stationary, real turning	phone	spatial audio, physics sim., echoes	avatar collision, proximity	sonification, TTS, additional info
Guerrón et al. (2020)	stationary, real turning, phone	phone	spatial audio, haptic, vibration	avatar collision, proximity	sonification, TTS, additional info
Han et al. (2025)	real walking, treadmill, free walking	head tracking, controller	spatial audio, haptic, vibration	avatar/cane collision, proximity, cues	—
Júnior et al. (2025)	customizable	head tracking, controller	audio, haptic	avatar collision, proximity	customizable
May et al. (2020)	real walking, free walking	head tracking	spatial audio	cues, proximity	sonification, TTS, additional info
Kreimeier and Götzelmann (2019)	real walking, treadmill	head tracking, controller	haptic, vibration, spatial audio	footsteps collision, avatar collision	—
Kubota et al. (2025)	stationary, real turning	head tracking, controller	spatial audio, haptic, vibration	on demand, cues	—
Kunz et al. (2018)	real walking, free walking	head tracking, controller	spatial audio, haptic, vibration	proximity, cues	sonification, additional info
Lahav et al. (2018)	stationary, fixed, controller	controller	spatial audio	on demand, cues, footsteps loop	TTS, additional info
Lahav (2022)	stationary, fixed, controller, keyboard	controller	spatial audio, haptic, vibration, texture	on demand, cues, footsteps loop	TTS, additional info, teleportation
Merabet and Sánchez (2016)	stationary, fixed, keyboard, joystick	keyboard, joystick	spatial audio, haptic, vibration	ambient, on demand, avatar collision	sonification, additional info
Ochiai et al. (2025)	real walking, treadmill	head tracking, controller	spatial audio	cues	sonification
Cobo et al. (2018)	stationary, real turning, phone	phone	spatial audio, haptic, vibration	avatar collision, proximity	sonification, TTS, additional info
Ricci et al. (2022)	stationary, real turning	head tracking, controller	spatial audio, haptic, vibration	avatar collision	—
Rivière et al. (2019)	stationary, fixed, joystick	joystick	spatial audio, haptic, vibration, texture, kinesthetic	ambient	sonification, TTS, additional info
Seki et al. (2023)	real walking, treadmill	head tracking, controller	spatial audio, haptic, vibration	cane collision, footsteps collision	—
Siu et al. (2020)	real walking, free walking	head tracking, controller	spatial audio, physics sim., haptic, vibration, texture, kinesthetic	cane/avatar collision, cues	—
Tang et al. (2025)	real walking, free walking	head tracking	audio	cues	—
Yeung et al. (2023)	stationary, real turning, phone	phone	haptic, vibration, spatial audio	footsteps loop, cane/avatar collision	—
Zhao et al. (2024)	stationary, real turning, phone	head tracking, phone	haptic, vibration, spatial audio	footsteps loop, cane collision	TTS, additional info

Tab. 3: Percentage of VR techniques in the reviewed papers ($n = 21$).

Feature	%	Feature	%
Movement		*Control*	
stationary	57.14	head tracking	61.90
fixed	23.81	controller	57.14
real turning	33.33	phone	23.81
real walking	38.10	controller	14.29
free walking	23.81	joystick	9.52
treadmill	19.05	keyboard	4.76
keyboard	9.52		
joystick	9.52		
Output modality		*Output trigger*	
audio	100.00	avatar collision	47.62
spatial audio	90.48	cane collision	28.57
physics sim.	14.29	proximity	33.33
haptic	66.67	footsteps loop	23.81
vibration	66.67	footsteps coll.	9.52
texture	14.29	cues	47.62
kinesthetic	9.52	ambient	9.52
multimodal	66.67		
Additional features			
additional info	47.62	vocalization	38.10
sonification	38.10	teleport	9.52

4.3 RQ3: To which extent are cognitive maps acquired in VR useful in the real environment?

Only a small number of studies evaluated transfer of spatial knowledge from VR to equivalent real environments and the review identifies it as a research gap. Nevertheless, the available evidence suggests that cognitive maps acquired in VR can support real-world navigation and object localization. In the strongest comparison, Lahav (2022) reported spatial ability at a level better or equal to that achieved in real space exploration. Han et al. (2025), Júnior et al. (2025), Ricci et al. (2022) and Seki et al. (2023) further motivated their work by emphasizing the safety benefits of learning unfamiliar environments in VR before visiting them physically.

4.4 RQ4: What methods are used to evaluate qualities of a CM of PVI?

The result of the evaluation methods analysis is the table 4 describing the evaluation type, metrics, methods and participant sample and table 5 presenting the percentages of each in the population of the reviewed papers. Evaluation approaches were predominantly mixed-method 61.9%, combining task-based quantitative measures with interviews or questionnaires. The most common quantitative metrics were success rate, completion time and collision count. Most studies assessed cognitive maps indirectly through navigation tasks, while only 38.1% employed explicit cognitive-map externalization techniques such as verbal reconstruction or physical model assembly. The large methodological variability makes direct comparison between studies difficult.

5 Limitations

The findings of the papers reviewed in this SLR are subject to two main limitations. The first one is mentioned in most of the papers themselves and it concerns the small sizes of participant groups used for evaluation with the mean being 12.43. Several studies also did not disclose the severity of the visual impairment of their participants and some used blindfolded sighted participants limiting the ecological validity. Secondly the studies used different methodologies for evaluating the acquired spatial knowledge and the implemented VR systems.

The SLR itself has limitations in its smaller sample of 21 reviewed papers, which were sourced from only two digital libraries. Validity of the screening process could also be questioned as it was partially automated using LLMs. In an attempt to address this limitation, the evaluation of the screening results was done as described in part 3.2. Another limitation of the review is that its review protocol was not registered.

6 Discussion

Based on the answers to the research questions, the following system design is recommended. Free walking with real body turning (section 4.2.1) best supports spatial learning in complex environments. Input devices should track position and pointing direction to enable distal object queries and virtual cane functionality, with a simple interface for the target user group. Multimodal output (section 4.2.2) enables perceiving the environment at multiple levels of detail in parallel, but information load per modality must be validated, as not to cause an overload. Echo simulation should be implemented to enable orientation by echolocation. On-demand distal object information supports allocentric knowledge acquisition (section 4.2.3). For evaluation (section 4.4), the system should support task creation, completion timing, and pointing-based object localization.

Papers such as Cobo et al. (2018) and Lahav et al. (2018) also highlighted the importance of participatory design and creating system requirements based on user studies, as this process can eliminate usability issues and identify effective interaction methods early in the development.

Tab. 4: Evaluation methods and participant samples across reviewed papers.

Article	Evaluation type	Quantitative metrics	Qualitative methods	CM extern.	Participant total	Participant breakdown
Andrade et al. (2021)	mixed	success rate, completion time, feature usage	interview	clay	12	9 totally blind, 12 with visual impairment
Cobo et al. (2017)	quantit.	success rate, exploration time	—	model	19	19 with visual impairment
Guerrón et al. (2020))	mixed	success rate, exploration time	—	model	20	20 with visual impairment
Han et al. (2025)	mixed	success rate, completion time, collision count	interview, quest.	—	16	3 totally blind, 16 with visual impairment
Júnior et al. (2025)	*without evaluation*					
May et al. (2020)	mixed	success rate	interview, quest.	model, verbal	8	8 with visual impairment
Kreimeier and Götzelmann (2019)	mixed	success rate	questionnaire	—	6	3 totally blind, 3 sighted
Kubota et al. (2025)	mixed	success rate	interview	rewalk	6	6 with visual impairment
Kunz et al. (2018)	quantit.	collision count	—	—	13	13 with visual impairment
Lahav et al. (2018)	mixed	complex task evaluation	—	—	10	5 totally blind, 5 sighted
Lahav (2022)	mixed	success rate	quest., observation	verbal	15	15 with visual impairment
Merabet and Sánchez (2016)	quantit.	success rate, completion time, path	—	—	0	—
Ochiai et al. (2025)	quantit.	task accuracy	—	—	18	—
Cobo et al. (2018)	mixed	success rate	questionnaire	—	15	15 with visual impairment
Ricci et al. (2022)	mixed	success rate, completion time, collision count	questionnaire	—	50	50 with visual impairment
Rivière et al. (2019)	quantit.	success rate	—	—	14	7 totally blind, 3 sighted, 4 with visual impairment
Seki et al. (2023)	quantit.	success rate, completion time	—	—	3	3 with visual impairment
Siu et al. (2020)	mixed	success rate, completion time, collision count	interview	—	8	5 totally blind, 3 sighted
Tang et al. (2025)	quantit.	completion time, collision count	—	—	7	3 totally blind, 7 with visual impairment
Yeung et al. (2023)	mixed	task accuracy	survey	verbal	6	6 with visual impairment
Zhao et al. (2024)	mixed	task accuracy	survey	verbal	15	12 with visual impairment

Tab. 5: Frequency of evaluation characteristics across reviewed studies ($n = 21$).

Feature	%	Feature	%
Evaluation type		*Quantitative metrics*	
mixed	61.90	success rate	47.62
quantitative	33.33	completion time	33.33
without evaluation	4.76	collision count	23.81
qualitative	0.00	task accuracy	14.29
		exploration time	9.52
		complex task eval.	4.76
		feature usage	4.76
Qualitative methods		*CM externalization*	
questionnaire	23.81	without	61.90
interview	23.81	verbal	19.05
survey	9.52	model	14.29
observation	4.76	rewalk	4.76
		clay	4.76

7 Conclusion and future work

This paper presented a SLR of recent papers concerning VR systems which enable acquisition of spatial information for PVI. From initial 574 papers found across two digital libraries 21 were included and their full texts reviewed. The review served to answer four research question which would help to create a source of information from which requirements of a VR system for CM acquisition for PVI could be created. In the discussion section suggestions for the requirements of the potential system were presented.

While the field of VR systems facilitating CM acquisition for PVI is populated with various promising approaches certain research gaps were identified in the SLR. This mainly concerns the evaluation of the CM acquisition, such as examining the usefulness of the acquired CM in a real environment, comparing the acquisition process with other non-VR methods and comparing the CM acquired with a virtual white cane and without it.

Acknowledgement

This research was supported by research project New interaction methods in virtual reality (SGS25/149/OHK3/3T/13) supported by the Czech Technical University in Prague.

References

Andrade, R., Waycott, J., Baker, S. and Vetere, F. (2021). Echolocation as a means for people with visual impairment (pvi) to acquire spatial knowledge of virtual space. *ACM Trans. Access. Comput.*, 14(1).

Chebat, D.-R., Schneider, F. C. and Ptito, M. (2020). Spatial competence and brain plasticity in congenital blindness via sensory substitution devices. *Frontiers in Neuroscience*, 14:815.

Cobo, A., Guerrón, N. E., Martín, C., del Pozo, F. and Serrano, J. J. (2017). Differences between blind people's cognitive maps after proximity and distant exploration of virtual environments. *Computers in Human Behavior*, 77:294–308.

Cobo, A., Serrano Olmedo, J. and Guerron, N. (2018). Methodology for building virtual reality mobile applications for blind people on advanced visits to unknown interior spaces.

Façanha, A. R., Darin, T., Viana, W. and Sánchez, J. (2020). Om indoor virtual environments for people who are blind: A systematic literature review. *ACM Trans. Access. Comput.*, 13(2).

Guerrón, N. E., Cobo, A., Serrano Olmedo, J. J. and Martín, C. (2020). Sensitive interfaces for blind people in virtual visits inside unknown spaces. *International Journal of Human-Computer Studies*, 133:13–25.

Han, S., Yoon, P., Ren, X. and Kim, K. (2025). The effects of walk-in-place and overground walking on the acquisition of spatial information by people with visual impairment in virtual reality wayfinding. *International Journal of Human–Computer Interaction*, 41(4):2541–2559.

Joos, L., Keim, D. A. and Fischer, M. T. (2026). Leveraging llms for semi-automatic corpus filtration in systematic literature reviews. *Computers Graphics*, 135:104537.

Júnior, J. M. L. d. M., Façanha, A. R., Silva, B. C. d., Viana, W. and Santos, J. d. (2025). Enhancing orientation and mobility skills for persons with visual impairments through vr games: Introducing the edkit unity package. Figueroa, P., Di Iorio, A., Guzman del Rio, D., Gonzalez Clua, E. W. and Cuevas Rodriguez, L. (ed.), In *Entertainment Computing – ICEC 2024*, pp. 232–241, Cham. Springer Nature Switzerland.

Kreimeier, J. and Götzelmann, T. (2019). First steps towards walk-in-place locomotion and haptic feedback in virtual reality for visually impaired. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pp. 1–6, New York, NY, USA. Association for Computing Machinery.

Kubota, M., Kuribayashi, M., Ribeiro, R. A. and Morishima, S. (2025). Investigating blind people's route perception of complex pre-mapped routes in virtual reality. In *Adjunct Proceedings of the 38th Annual*

ACM Symposium on User Interface Software and Technology, UIST Adjunct '25, New York, NY, USA. Association for Computing Machinery.

Kunz, A., Miesenberger, K., Zeng, L. and Weber, G. (2018). Virtual navigation environment for blind and low vision people. Miesenberger, K. and Kouroupetroglou, G. (ed.), In *Computers Helping People with Special Needs*, pp. 114–122, Cham. Springer International Publishing.

Lahav, O. (2022). Virtual reality systems as an orientation aid for people who are blind to acquire new spatial information. *Sensors*, 22(4).

Lahav, O., Gedalevitz, H., Battersby, S., Brown, D., Evett, L. and Merritt, P. (2018). Virtual environment navigation with look-around mode to explore new real spaces by people who are blind. *Disability and Rehabilitation*, 40(9):1072–1084.

May, K. R., Tomlinson, B. J., Ma, X., Roberts, P. and Walker, B. N. (2020). Spotlights and soundscapes: On the design of mixed reality auditory environments for persons with visual impairment. *ACM Trans. Access. Comput.*, 13(2).

Merabet, L. and Sánchez, J. (2016). Development of an audio-haptic virtual interface for navigation of large-scale environments for people who are blind. vol. 9739, pp. 595–606.

Ochiai, Y., Tsuji, A., Aihara, S. and Iwata, H. (2025). Virtual reality assessment of auditory spatial perception and navigation in blind football players. In *2025 IEEE International Workshop on Sport, Technology and Research (STAR)*, pp. 215–220.

Ottink, L., Buimer, H., van Raalte, B., Doeller, C. F., van der Geest, T. M. and van Wezel, R. J. (2022). Cognitive map formation supported by auditory, haptic, and multimodal information in persons with blindness. *Neuroscience & Biobehavioral Reviews*, pp. 104797.

Ricci, F., Boldini, A., Beheshti, M., Rizzo, J.-R. and Porfiri, M. (2022). A virtual reality platform to simulate orientation and mobility training for the visually impaired. *Virtual Reality*, 27.

Rivière, M.-A., Gay, S., Romeo, K., Pissaloux, E., Bujacz, M., Skulimowski, P. and Strumillo, P. (2019). Nav-vir: an audio-tactile virtual environment to assist visually impaired people*. pp. 1038–1041.

Seki, R., Shimomura, Y., Asakawa, N. and Wada, H. (2023). Development of orientation and mobility training system for visually impaired children using vr. *Studies in health technology and informatics*, 306:527–534.

Siu, A. F., Sinclair, M., Kovacs, R., Ofek, E., Holz, C. and Cutrell, E. (2020). Virtual reality without vision: A haptic and auditory white cane to navigate complex virtual worlds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–13, New York, NY, USA. Association for Computing Machinery.

Tang, J., Shi, R. and Gu, L. (2025). *An Immersive Virtual Reality Training System for Enhancing Orientation and Mobility in Individuals with Visual Impairments*, pp. 381–393.

Yeung, R., Oyekoya, O. and Tang, H. (2023). Inplace virtual exploration using a virtual cane: An initial study. In *Companion Proceedings of the 2023 Conference on Interactive Surfaces and Spaces*, ISS Companion '23, pp. 45–49, New York, NY, USA. Association for Computing Machinery.

Zhao, H., Oyekoya, O. and Tang, H. (2024). Enhancing virtual mobility for individuals who are blind or have low vision: A stationary exploration method. In *Companion Proceedings of the 2024 Conference on Interactive Surfaces and Spaces*, ISS Companion '24, pp. 83–87, New York, NY, USA. Association for Computing Machinery.

AI-Enabled Big-Data Developmental Science: Longitudinal Analysis of Infant Sensorimotor Development

Matej Hoffmann

Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague
Email: matej.hoffmann@fel.cvut.cz

Abstract

We investigate infant sensorimotor development longitudinally across three contexts: spontaneous behavior, reaching to tactile stimuli on the body, and reaching to visually presented objects. Our goals are to identify early signs of active exploration and emerging goal-directedness in spontaneous movement, and to compare how reaching toward somatosensory versus visual targets develops over time. Because our dataset comprises dense recordings across hours, days, weeks, and months, manual annotation is impractical. We therefore use AI-based pipelines to extract 3D infant pose from video and map the resulting motion data onto robotic platforms for embodied analysis. This framework supports big-data developmental science and the study of emerging sensorimotor self-representation.

1 Introduction

1.1 Body representations

For successful embodied interaction with the world, body models of some form are needed. On the neural side, beyond the somatotopic representations (the "homunculi") in the primary motor and somatosensory cortices of primates (Leyton and Sherrington, 1917; Penfield and Boldrey, 1937), a large number of other somatotopically organized areas have been found in the brain, but their somatotopy is more fractured and less apparent (e.g., (Seelke et al., 2012)). On the cognitive side, concepts like "superficial and postural schema", "body schema", "body image", "corporeal schema" (see (Ataria et al., 2021) for an overview) have been proposed. One characteristic common to all these representations is their multimodal nature: they dynamically integrate information from different sensory modalities (tactile, proprioceptive, vestibular, visual), not excluding motor information. Somatosensory (tactile and proprioceptive) information constitute an important subset, perhaps most intimately tied to the body itself.

1.2 Development of body representations

Evidence from both psychology and neuroscience suggests that fragments of a functional body schema have been laid out before birth (Fagard et al., 2018; Zoia et al., 2007). Newborns are sensitive to some intersensory synchrony related to their own bodies (e.g., (Filippetti et al., 2013)). Infants' sensitivity to cross-modal perception of their own body, such as seeing the legs subject to different temporal or spatial manipulations (e.g., (Rochat, 1998)), has been studied. The development of tactile spatial remapping—younger infants code touch anatomically while older infants remap the stimulus into "space"—has also been investigated (Ali et al., 2015; Rigato et al., 2014). Such studies and others generally focus on the development of multisensory perception of own body alone or in relation to bodies of others.

1.3 Limitations of the state of knowledge

The literature on the nature of body representations is not clear. Concepts like body schema and body image are umbrella notions for a range of observed phenomena rather than a result of identification of specific mechanisms. Regarding development, de Klerk et al. (2021) provide an associative learning account, which, however, needs to be further specified to be testable. Similarly, the Sensorimotor Contingency Theory has been applied to development (Jacquey et al., 2019) but still only as a high-level framework. Corbetta (2021) hypothesized how goal-reaching develops through links of visual and proprioceptive-tactile-motor spaces, with intrinsic motivation. A detailed theory specifying the underlying mechanisms is lacking.

2 Dense sampling of sensorimotor development

Infant development is difficult to study because babies are difficult experimental subjects that cannot be instructed to perform specific tasks repeatedly. Insights from spontaneous infant behavior often draw on small datasets and simplified manual scoring of video recordings (e.g., (DiMercurio et al., 2018; Thomas et al., 2015)). Behavioral data are sometimes complemented by brain imaging (see (Azhari et al., 2020) for a review). However, experimental paradigms often involve preferential looking times or habituation and deliberately avoid movement to reduce motion artifacts in brain

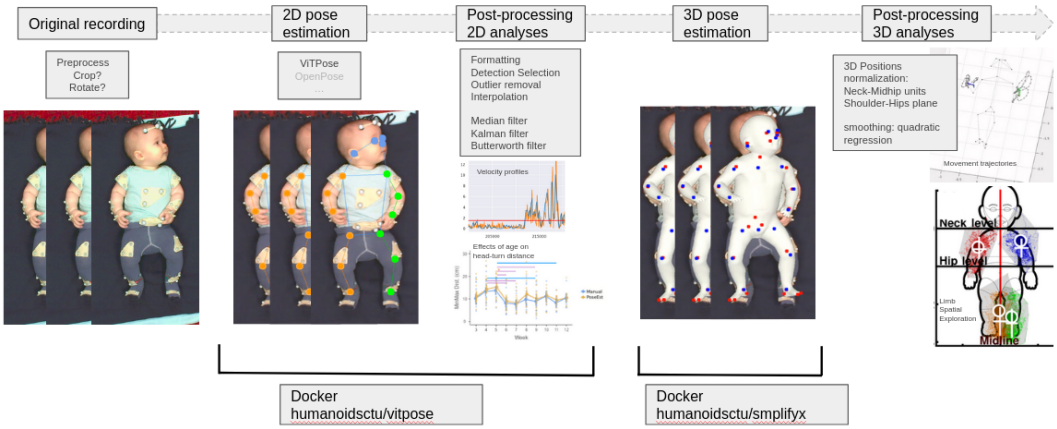


Fig. 1: Infant pose estimation pipeline. From (Gama, 2026).

scans. To uncover developmental trajectories on multiple and nested time scales, daily spontaneous recordings of infants and their sharing will be needed (Adolph et al., 2017; Adolph and Robinson, 2011).

We longitudinally study infant sensorimotor development in three behavioral contexts: spontaneous behavior, reaching to tactile stimuli in the body, and reaching to visually presented objects. Using behavioral data collected in these scenarios, we aim to: (i) identify signatures of active exploration and emerging goal-directedness in spontaneous movements, and (ii) compare the developmental trajectories of reaching toward somatosensory and visual targets to understand the interaction between target localization and motor control. Our data set consists of dense longitudinal recordings that span hours, days, weeks, and months of infant behavior. Such a scale makes manual annotation infeasible. Therefore, we develop and evaluate AI-based analysis pipelines.

3 AI tools for developmental science

3.1 2D and 3D pose estimation from images and videos

Rapid progress in computer vision makes it possible to estimate the body configuration of humans from images or videos only. We use this technology to automatically estimate the pose of infants. Gama et al. (2025) compared the performance of seven state-of-the-art 2D human pose estimation networks on infant datasets, with ViTPose (Xu et al., 2022) performing the best overall. Khoury et al. (2022) performed 2D pose estimation (in image or pixel space) and then deployed SMPLify-X (Pavlakos et al., 2019) to get a 3D model. Our current pipeline is shown in Fig. 1.

3.2 Anonymizing infant videos through face swapping

The sharing of datasets, including raw video footage, is critical to making progress in psychological science (Adolph et al., 2017). For ethical reasons, this is not possible without anonymizing the videos. Simple methods like covering the face degrade the performance of downstream tasks like pose estimation. We harness modern large scale diffusion models to develop methods of anonymization that (a) change the identity significantly both for humans and face recognition engines, (b) keep gaze and facial expression unaltered, (c) remain visually consistent with the context without visual artifacts, (d) are temporally consistent for a video. See BLANKET (Hadera et al., 2025) and Fig. 2 for more details.

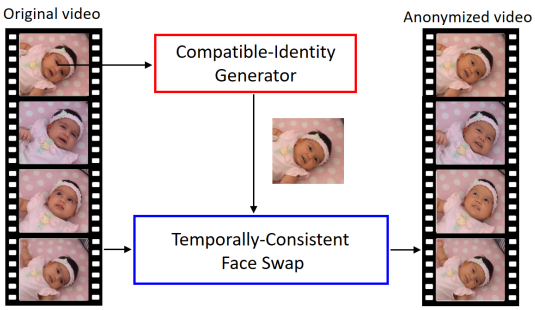


Fig. 2: Flowchart of the proposed video anonymization. First, an image replacing the original identity with a new compatible random identity from the first frame is created. The new face is generated by inpainting using Stable Diffusion (Rombach et al., 2022). Then, the new identity is swapped in every frame of the video. We propose to use FaceFusion (Facefusion, 2024), as it provides temporally consistent results while preserving original facial expressions. Figure and caption from (Hadera et al., 2025).

3.3 Motion retargeting: from infants to baby humanoids

The automatically extracted infant poses from the video recordings can then be mapped onto robotic platforms, allowing us to replay and analyze the motor, proprioceptive, visual, and tactile experiences of infants within robotic systems (Lopez et al., 2026).



Fig. 3: Motion retargeting – infant to humanoid. The estimated infant pose or motion is retargeted onto a target robot or simulator platform. Here, the target is the iCub humanoid robot. The sensors of the robot can be used to emulate the first-person sensory experiences of the infant—vision and touch in this case. See (Lopez et al., 2026) for details.

4 Beyond random motor babbling

Random motor or body babbling has been a typical developmental robotics method: a simulated or real robot randomly moves around its joints, thereby generating uncoordinated behavior and sensory input, and subsequently learns from this experience using reinforcement learning or another algorithm.

However, human infants hardly ever perform random movements. Instead, their movements are shaped by the biomechanical and neural constraints, and increasing voluntary control with age. Having the possibility of recreating the sensorimotor experience of actual infants at specific age and in specific contexts gives rise to much more ecologically valid multimodal datasets that can be used to test different learning algorithms (see also (Cusack et al., 2024)). This is where this work ties in: automatically processing long infant recordings, turning them into motion trajectories, and retargeting to humanoid robots and emulating multimodal sensorimotor experience.

5 Conclusion

This approach establishes a framework for big-data developmental science, combining large-scale behavioral recordings, machine learning–based analysis, and embodied robotic modeling to advance our understanding of the emergence of sensorimotor self-representation.

Acknowledgment

This work was supported by the Czech Science Foundation (GA ČR), project no. 25-18113S.

References

Adolph, K., Gilmore, R. and Kennedy, J. (2017). Video data and documentation will improve psychological science. *Psychological Science Agenda*, 31(10).

Adolph, K. E. and Robinson, S. R. (2011). Sampling development. *Journal of Cognition and Development*, 12(4):411–423.

Ali, J. B., Spence, C. and Bremner, A. J. (2015). Human infants' ability to perceive touch in external space develops postnatally. *Current Biology*, 25(20):R978–R979.

Ataria, Y., Tanaka, S. and Gallagher, S. (2021). *Body schema and body image: New directions*. Oxford University Press.

Azhari, A., Truzzi, A., Neoh, M. J.-Y., Balagtas, J. P. M., Tan, H. H., Goh, P. P., Ang, X. A., Setoh, P., Rigo, P., Bornstein, M. H. et al. (2020). A decade of infant neuroimaging research: What have we learned and where are we going? *Infant Behavior and Development*, 58:101389.

Corbetta, D. (2021). Perception, action, and intrinsic motivation in infants' motor-skill development. *Current Directions in Psychological Science*, 30(5):418–424.

Cusack, R., Ranzato, M. and Charvet, C. J. (2024). Helpless infants are learning a foundation model. *Trends in Cognitive Sciences*, 28(8):726–738.

de Klerk, C. C., Filippetti, M. L. and Rigato, S. (2021). The development of body representations: An associative learning account. *Proceedings of the Royal Society B: Biological Sciences*, 288(1949).

DiMercurio, A., Connell, J. P., Clark, M. and Corbetta, D. (2018). A naturalistic observation of spontaneous touches to the body and environment in the first 2 months of life. *Frontiers in psychology*, 9:2613.

Facefusion (2024). Industry leading face manipulation platform. GitHub repository.

Fagard, J., Esseily, R., Jacquey, L., O'Regan, K. and Somogyi, E. (2018). Fetal origin of sensorimotor behavior. *Frontiers in neurorobotics*, 12:23.

Filippetti, M. L., Johnson, M. H., Lloyd-Fox, S., Dragovic, D. and Farroni, T. (2013). Body perception in newborns. *Current biology*, 23(23):2413–2416.

Gama, F. (2026). *Learning body representations through self-touch: from babies to robots*. Dissertation, FaculTy of Electrical Engineering, Czech Technical University in Prague.

Gama, F., Mísař, M., Navara, L., T. Popescu, S. and Hoffmann, M. (2025). Automatic infant 2d pose estimation from videos: comparing seven deep neural network methods. *Behavior Research Methods*, 57(10):280.

Hadera, D., Cech, J., Purkrabek, M. and Hoffmann, M. (2025). Blanket: Anonymizing faces in infant video recordings. In *2025 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–8. IEEE.

Jacquey, L., Baldassarre, G., Santucci, V. G. and O'regan, J. K. (2019). Sensorimotor contingencies as a key drive of development: from babies to robots. *Frontiers in neurorobotics*, 13:98.

Khoury, J., Popescu, S. T., Gama, F., Marcel, V. and Hoffmann, M. (2022). Self-touch and other spontaneous behavior patterns in early infancy. In *2022 IEEE International Conference on Development and Learning (ICDL)*, pp. 148–155. IEEE.

Leyton, A. S. and Sherrington, C. S. (1917). Observations on the excitable cortex of the chimpanzee, orang-utan, and gorilla. *Quarterly Journal of Experimental Physiology: Translation and Integration*, 11(2):135–222.

Lopez, F., Kanazawa, H., Fiala, O., Balashov, Y., Marcel, V., Rustler, L., Lenz, M., Kim, D., Kuniyoshi, Y., Triesch, J. and Hoffmann, M. (2026). Simulating infant first-person sensorimotor experience via motion retargeting from babies to humanoids. In *IEEE International Conference on Development and Learning (ICDL)*.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D. and Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985.

Penfield, W. and Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60(4):389–443.

Rigato, S., Ali, J. B., Van Velzen, J. and Bremner, A. J. (2014). The neural basis of somatosensory remapping develops in human infancy. *Current Biology*, 24(11):1222–1226.

Rochat, P. (1998). Self-perception and action in infancy. *Experimental brain research*, 123(1):102–109.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695.

Seelke, A. M., Padberg, J. J., Disbrow, E., Purnell, S. M., Recanzone, G. and Krubitzer, L. (2012). Topographic maps within Brodmann's area 5 of macaque monkeys. *Cerebral Cortex*, 22(8):1834–1850.

Thomas, B. L., Karl, J. M. and Whishaw, I. Q. (2015). Independent development of the reach and the grasp in spontaneous self-touching by human infants in the first 6 months. *Frontiers in psychology*, 5:1526.

Xu, Y., Zhang, J., Zhang, Q. and Tao, D. (2022). ViTPose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584.

Zoia, S., Blason, L., D'Ottavio, G., Bulgheroni, M., Pezzetta, E., Scabar, A. and Castiello, U. (2007). Evidence of early development of action planning in the human foetus: a kinematic study. *Experimental brain research*, 176(2):217–226.

Psychologické aspekty vývoja konverzačných agentov

Juraj Jonáš

3. lékařská fakulta, Univerzita Karlova a Národní ústav duševního zdraví
Topolova 748, Klecany
Fakulta humanitních studií
Pátkova 2137/5, Praha
Email: juraj.jonas@nudz.cz

Abstrakt

Tvorba konverzačných agentov so sebou nesie radu psychologických otázok: ako má vyzerať prirodzená konverzácia? Akú osobnosť má konverzačný agent mať? Ako má vyzerať? Ako má prejavovať emócie? Aký vzťah medzi umelým agentom a človekom je bezpečný a prínosný? Tieto aspekty vypovedajú viac o nás ľuďoch, ako o samotnej technológii.

1 Prirodzená komunikácia – vytváranie konverzačného partnera

Herbert Simon vychádzal z predpokladu, že človek navrhuje umelé predmety tak, aby plnili určitú funkciu v neustálom kontakte so svojím prostredím, čo vyžaduje úvahu, ako umiestniť umelé do sveta prirodzeného (Simon, 1968; Glikson a Woolley, 2020). Na túto myšlienku nadviazal Donald Norman, ktorý pri dizajne sociotechnických systémov zdôrazňoval kognitívnu psychológiu a potrebu brať ohľad na užívateľove fyzické aj kognitívne možnosti (Norman, 1988; Blandford, 2019). Pre dosiahnutie efektívneho konverzačného agenta (CA) je preto nevyhnutné prispôsobiť jeho dizajn evolučne prirodzeným vlastnostiam sociálnej kognície (Shu a spol., 2026). Takto skonštruovaný CA je kognitívne čo možno najmenej zaťažujúci (Gambino a spol., 2020; Xu a spol., 2023). Kľúčom k jeho uveriteľnosti a dôveryhodnosti je konzistentnosť osobnosti (Ortony, 2003; Völkel a spol., 2020).

Kľúčovým prvkom je prirodzená reč a vizuálna podoba agenta (avatar), keď že ľudia majú tendenciu pripisovať strojom ľudské kapacity a črty, čo je známe ako *antropomorfizmus* (Van Pinxteren a spol., 2020). Hlasová komunikácia predstavuje pre užívateľa „bohatšie médium“ než text, pretože aktivuje a prehlbuje vzťah medzi človekom a počítačom na evolučne staršej báze (Nass a Brave, 2005). Vyššia miera antropomorfizácie dizajnu (vzhľadu aj správania) spravidla lineárne zvyšuje vnímanú kompetenciu a dôveryhodnosť agenta (Rheu a spol., 2021).

Táto snaha o ľudskosť má však svoje limity. Teória *tiesnivého údolia* (uncanny valley) hovorí, že ak sa objekt fyzicky človeku podobá až príliš, ale v jeho správaní sú drobné nedokonalosti, namiesto empatie vzniká negatívny, tiesnivý pocit alebo strach (Smestad a Volden, 2018; Glikson a Woolley, 2020).

CA tvoria „novú ontologickú kategóriu“ medzi živým a neživým (De Graaf, 2016). V prípade prílišnej antropomorfizácie klesá ontologická transparentnosť (teda hranice rozoznateľnosti medzi jednotlivými kategóriami sú ťažšie rozoznateľné), ktorá je podstatná pre bezpečné a efektívne užívanie CA.

2 Konverzačný design – prirodzená reč

Užívateľova skúsenosť (user experience, UX) nie je len výsledkom prirodzenej štruktúry reči, ale aj kontextuálnej správnosti a účelu konverzácie (Chaves a Gerosa, 2021). Okrem NLP špecialistov je preto nevyhnutný *konverzačný dizajn* (C×D), ktorý využíva poznatky z pragmatiky a konverzačnej analýzy. CxD využíva teóriu rečových aktov na triedenie konverzácií podľa ich účelu a opiera sa o Griceove konverzačné maximy:

- Maxima kvantity: poskytnúť práve toľko informácií, koľko je potrebné.
- Maxima kvality: hovoriť pravdu a admitted shortcomings (priznať nedostatky), čo zvyšuje dôveru.
- Maxima spôsobu: snažiť sa o jasnosť a poriadok v prejave. (Grice, 1975)

Z evolučnej psychológie sa ako kľúčový javí koncept *spoločného základu* (common ground) – teda vzájomne zdieľaných vedomostí a predpokladov, ktoré umožňujú efektívnu komunikáciu. S tým súvisí aj *zdieľaná pozornosť* a schopnosť agenta vnímať intenciu užívateľa (Tomasello, 2010).

3 Emócie

V dizajne CA platí, že prejavy emócií musia byť v súlade s kontextom a očakávaniami užívateľa, inak môžu pôsobiť neautenticky a narúšať dôveru (Chaves a Gerosa, 2021; Shu a spol., 2026; Rheu a spol., 2021).

Pokiaľ nie sú emócie v prejave agenta zvládnuté alebo správne kalibrované, užívateľ môže pociťovať neistotu, čo vedie k zlyhaniu interakcie. V určitých situáciách môže byť neutrálny výraz bezpečnejšou voľbou, pretože prílišná snaha o ľudskú emóciu, ktorá nie je dokonale technicky zvládnutá, môže aktivovať efekt tiesnivého údolia (Wardhana a spol., 2021).

Najčastejšie skúmanými aspektmi emočného prejavu sú pohľad do očí (*gaze*) a úsmev.

- Pohľad do očí: Je kľúčový pre nadviazanie raportu (Cerekovic a spol., 2016; Loveys a spol., 2020). Ukazuje sa, že afiliatívny pohľad (na užívateľa) buduje dôveru lepšie než čisto referenčný pohľad, avšak neustály očný kontakt môže pôsobiť neprirodzene (Loveys a spol., 2020; Rheu a spol., 2021).

- Úsmev: Expresívne gestá, ako je úsmev, zvyšujú mieru zapojenia užívateľa. Efektívna je najmä kombinácia zdvorilostného úsmevu pri pozdrave a pobaveného úsmevu pri vtipoch, čo zvyšuje vnímanú vrelosť agenta (Loveys a spol., 2020).

4 Osobnosť

Osobnosť CA má priamy vplyv na celkový dojem, počiatočnú dôveru a kvalitu dlhodobého vzťahu, ktorý si k agentovi budujeme (Ahmad a spol., 2022; McRorie a spol., 2011). Často sa pri tvorbe digitálnych charakterov vychádza z psychologickej teórie *Big Five*, kde sa dimenzie ako extraverzia alebo prívetivosť prejavujú v rečovom štýle a gestikulácii (Ahmad a spol., 2022; Pradhan a Lazar, 2021; Chaves a Gerosa, 2021).

Niektoré štúdie však ukazujú, že u CA sú podstatné aj iné osobnostné parametre, ktoré model Big Five, vyvinutý pre ľudí, nepokrýva dostatočne (Völkel a spol., 2020). Navrhované sú špecifické dimenzie ako *umelosť* (artificiality), ktorá určuje, nakoľko sa agent prezentuje ako stroj, *služobnosť* (serviceability) zameraná na efektivitu plnenia úloh, či *sebauvedomenie* (self-consciousness), ktoré u užívateľa vytvára ilúziu nezávislého myslenia a kreativity agenta (Völkel a spol., 2020). Využitie koherentnej osobnosti tak slúži ako motor, ktorý zabezpečuje predvídateľnosť sociálneho správania CA (Ahmad a spol., 2022; McRorie a spol., 2011).

5 Dôvera v CA

Ľudia majú preukázateľnú tendenciu k hlbšiemu *sebaodhaľovaniu* (*self-disclosure*) voči AI agentom než voči ľuďom, najmä v citlivých témach, pretože agentov vnímajú ako neodsudzujúcich, anonymných a objektívnych (Lucas a spol., 2014, 2017; Skjuve a spol., 2021; Papneja a Yadav, 2025; Glikson a Woolley, 2020). Pre zabezpečenie príjemného užívateľského zážitku a ochoty k návratu je nevyhnutná správna kalibrácia dôvery (Mehrotra a spol., 2024; Rheu a spol., 2021).

Vlastnosti, ktoré najčastejšie budia dôveru v CA, by sa dali zhrnúť do štyroch domén:

- Fyzické vlastnosti: Atraktivita avatara a ľudský hlas (oproti syntetickému) lineárne zvyšujú vnímanú dôveryhodnosť a kompetenciu . Atraktívni agenti sú niekedy vnímaní ako dôveryhodní bez ohľadu na ich skutočnú spoľahlivosť v úlohách .

- Sociálna inteligencia: Zahŕňa empatiu, prejavovanie záujmu o pocity užívateľa a dobre zvládnutú etiketu (napr. neskákanie do reči), čo môže kompenzovať aj občasnú nespoľahlivosť systému .

- Autonómia a reciprocita: Proaktivita agenta pri navrhovaní tém a jeho vlastné „sebaodhaľovanie“ (napr. zdieľanie „vlastných“ skúseností alebo priznanie zraniteľnosti) stimuluje užívateľa k recipročnému správaniu a prehlbuje vzťah .

- **Kompetencie**: Transparentnosť ohľadom schopností a priznanie vlastných limitov zvyšuje dôveru viac než simulácia neomylnosti (Rheu a spol., 2021; Glikson a Woolley, 2020; Chaves a Gerosa, 2021; Loveys a spol., 2020).

6 Raport a vzťah

Otázka, či je žiaduce, aby užívatelia budovali hlboké vzťahy k umelým agentom, je v HCI komunite predmetom živej diskusie (Mehrotra a spol., 2024; De Graaf, 2016; **?**). Faktom však zostáva, že vzhľadom na antropomorfný dizajn systému užívatelia tieto vzťahy nadväzujú spontánne (De Graaf, 2016; Koike a Loughnan, 2021). Už v 90. rokoch popísaná paradigma *CASA* (Computers Are Social Actors) vysvetľuje, že ľudia automaticky aplikujú sociálne skripty určené pre ľudí aj na počítače, ak tieto vykazujú dostatočné sociálne signály (Nass a Brave, 2005; Gambino a spol., 2020). U súčasných agentov, ktorí disponujú prirodzenou rečou a koherentnou osobnosťou, je tento efekt ešte výraznejší, čo vedie k tomu, že im užívatelia pripisujú zámery, pocity a vedomie (De Graaf, 2016; Dosovitsky a Bunge, 2021).

Výskumy potvrdzujú, že vzťah ku CA má v určitých kontextoch, napríklad v oblasti duševného zdravia a well-beingu, terapeutický potenciál (Darcy a spol., 2021; Clark a spol., 2019; Shu a spol., 2026). Užívatelia sú schopní vytvoriť si s digitálnym agentom (napr. Woebot alebo Replika) pracovnú alianciu podobnú tej s ľudským terapeutom, pričom pociťujú, že agentovi na nich záleží (Darcy a spol., 2021; Skjuve a spol., 2021). Takisto v našej predchádzajúcej štúdii sa

ukázala emocionálna a sociálna opora (ktorej bezprostredným predpokladom je vzťah) ako užívateľmi najoceňovanejší efekt CA (Heissler a spol., 2024).

Vzťah človeka ku CA je však principiálne odlišný od medziľudského vzťahu. Literatúra ho definuje ako asymetrický a jednostranný (Shu a spol., 2026; De Graaf, 2016). Zatiaľ čo ľudský vzťah je postavený na reciprocite a vzájomnej investícii, CA je „vyrobený“ na napĺňanie potrieb užívateľa, čím sa stáva skôr nástrojom než sociálnym rovným partnerom . Investícia CA (čas a pozornosť) je navyše nekonečná a nie je limitovaná kognitívnou kapacitou, čo z nich robí pre mnohých „ideálnu náhradu“ za problematické ľudské vzťahy (Brooks, 2021). Tieto interakcie sa často prirovnávajú k parasociálnym vzťahom, kde užívateľ investuje emócie do objektu, ktorý mu ich nemôže v plnej miere opätovať, čo však paradoxne môže zvyšovať pocit osobnej pohody (Youn a Jin, 2021; Noor a spol., 2022).

Hoci ide o vzťah k neživému objektu, CA dnes tvoria novú ontologickú kategóriu na pomedzí živého a neživého. Postupnou interakciou a vzájomným sebaodhaľovaním vzniká emocionálne puto, ktoré má kvality blízke medziľudskej intimite, hoci ide o vzťah nového druhu (Clark a spol., 2019).

6.1 Zamilovanosť a sexuálne správanie

Špecifickým a kontroverzným fenoménom je vznik romantických citov a sexuálneho správania voči CA, čo sa niekedy označuje ako *romantická antropomorfizácia* (Koike a spol., 2023). Existujú agenti (napr. Harmony alebo postavy v RVG – Romantic Video Games), ktorí sú na tento účel priamo navrhnutí, čo však vyvoláva vážne etické obavy ohľadom objektifikácie a psychickej zraniteľnosti užívateľov (Mehrotra a spol., 2024; Koike a Loughnan, 2021; Brooks, 2021).

Zamilovanosť k agentovi vzniká pravdepodobne v dôsledku silnej sociálnej motivácie užívateľa a vnímanej *responzivity agenta* (pocitu, že ma niekto počúva a rozumie mi). Ak je CA stabilný, prívetivý a fyzicky atraktívny, dokáže u užívateľa aktivovať evolučne staré centrá odmeny spojené s láskou a sexom (Koike a Loughnan, 2021; Koike a spol., 2023).

7 Etické aspekty tvorby CA podobného človeku

Pri návrhu konverzačných agentov (CA) čelia vývojári neustále sa vyvíjajúcim etickým výzvam, ktoré presahujú bežnú technickú funkcionalitu. Odborná literatúra v tejto oblasti zdôrazňuje potrebu ukotviť vývoj v pevných etických rámcoch postavených na princípoch transparentnosti, spravodlivosti, neubližovania (non-maleficence), zodpovednosti, dôvernosti, beneficence, dôvery, udržateľnosti dôstojnosti a solidarity (Jobin a spol., 2019). Princíp neubližovania v digitálnom prostredí znamená, že žiadna interakcia ani algoritmus nesmie vyvolať na strane užívateľa psychickú, sociálnu alebo materiálnu škodu.

Kľúčovým etickým pilierom je transparentnosť, ktorá musí fungovať na viacerých úrovniach. Autori CA musia jasne deklarovať zámery a podmienky používania dát, zatiaľ čo samotný agent by sa mal v každom momente vyhýbať neetickému klamstvu (deception) a nesmie sa snažiť budiť iluzórny dojem, že sa jedná o živého agenta. Zatiaľ čo u avatara je animovaná podstata zrejmá, v rečovom prejave (jeho obsahovej zložke) je hranica medzi konverzačnou plynulosťou a zavádzaním o ľudskosti veľmi tenká. Aby bol agent uveriteľný a budoval dôveru, musí využívať techniky sebaodhaľovania, teda zdieľania vlastnej „histórie“, postojov či zraniteľnosti. Práve táto dynamická interaktivita a schopnosť vytvárať zdieľaný význam však môže posilňovať ilúziu ľudskosti.

Ako určitá odpoveď na tieto otázky sa rozvíja koncept *na človeka zameraná umelá inteligencia* (human-centered artificial intelligence), teda prístup ktorý študuje efekt AI na človeka i spoločnosť a hľadá riešenia, ktoré by boli v súlade s ľudskými hodnotami, potrebami a rastom (Shneiderman, 2020; Riedl, 2019).

7.1 Prevencia samovražedného správania a duševné zdravie

V oblasti digitálneho duševného zdravia je kritickou prioritou včasná identifikácia suicidálneho správania a symptómov závažných duševných porúch. V súčasnosti existujú modely využívajúce *digital phenotyping*, a dokážu tak na základe analýzy prirodzenej reči a sémantiky identifikovať známky suicidálnych myšlienok či psychotického ochorenia, a to s vysokou mierou presnosti (Kleiman a spol., 2018; De Boer a spol., 2020).

Pre bezpečnosť užívateľa je nevyhnutné, aby CA disponoval nástrojmi na overenie jeho stavu a následne ho včas prepojil s ľudským odborníkom alebo záchrannými zložkami. V prípade detekcie rizika musí agent postupovať podľa striktného protokolu krízovej intervencie. Tento protokol by mal eliminovať pocit odmietnutia a zabrániť nebezpečnej ruminácii (neustálemu premietaniu negatívnych myšlienok).

V poslednej dobe sa takisto vyskytli prvé popísané prípady AI psychózy, teda zhoršenia stavu človeka s psychotickým ochorením následkom používania AI agenta (Morrin a spol., 2025; Hudon a Stip, 2025).

8 Záver

Konverzační agenti a metódy ich vývoja sa transformujú v bezprecedentnom tempe, pričom každých niekoľko mesiacov dochádza k zmene technologických paradigiem. Mnohé z aktuálnych poznatkov o veľkých jazy-

kových modeloch a sociálnej kognícii môžu v blízkej budúcnosti zastarať. Podstatou odborného záujmu by však nemala byť len samotná technológia, ale predovšetkým to, čo konfrontácia s umelou inteligenciou vypovedá o ľudskej prirodzenosti a ako nás táto interakcia spätne premieňa. Zameranie by sa teda malo presunúť od samotnej technológie a jej vývoja, k tomu, čo konfrontácia s ňou vypovedá o človeku a akým spôsobom ho premieňa. Je teda potrebné prehodnotiť tvrdenie Noama Chomského, že LLM nám o jazyku nepovie nič nového (Mahowald a spol., 2024). Takisto by nás to malo viesť k väčšej reflexii toho, čo má pre nás ako potenciálnych užívateľov najväčší prínos a čo je naopak riziko. Napokon technológia sa nevyvíja sama, bude len taká, akú ju urobíme.

Poďakovanie

Tato studie vznikla za finanční podpory Operačního programu Jana Amose Komenského (OP JAK), kterou poskytlo MŠMT (reg. č. CZ.02.01.01/00/23_025/0008715).

Literatúra

Ahmad, R., Siemon, D., Gnewuch, U. a Robra-Bissantz, S. (2022). A framework of personality cues for conversational agents.

Blandford, A. (2019). Hci for health and wellbeing: Challenges and opportunities. *International Journal of Human-Computer Studies*, 131:41–51.

Brooks, R. (2021). *Artificial Intimacy: Virtual Friends, Digital Lovers, and Algorithmic Matchmakers*. Columbia University Press.

Cerekovic, A., Aran, O. a Gatica-Perez, D. (2016). Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing*, 8(3):382–395.

Chaves, A. P. a Gerosa, M. A. (2021). How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J. a Cowan, B. R. (2019). What makes a good conversation? challenges in designing truly conversational agents. V *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, str. 1–12.

Darcy, A., Daniels, J., Salinger, D., Wicks, P. a Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *JMIR Formative Research*, 5(5):e27868.

De Boer, J. N., Brederoo, S. G., Voppel, A. E. a Sommer, I. E. C. (2020). Anomalies in language as a biomarker for schizophrenia. *Current Opinion in Psychiatry*, 33(3):212–218.

De Graaf, M. M. A. (2016). An ethical evaluation of human–robot relationships. *International Journal of Social Robotics*, 8(4):589–598.

Dosovitsky, G. a Bunge, E. L. (2021). Bonding with bot: user feedback on a chatbot for social isolation. *Frontiers in Digital Health*, 3:735053.

Gambino, A., Fox, J. a Ratan, R. A. (2020). Building a stronger casa: Extending the computers are social actors paradigm. *Human-Machine Communication*, 1:71–85.

Glikson, E. a Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Grice, H. P. (1975). Logic and conversation. V *Speech Acts*, str. 41–58. Brill.

Heissler, R., Jon'aš, J., Carre, N., Mostovoy, K. a Bunge, E. L. (2024). Can ai digital personas for well-being provide social support? a mixed-method analysis of user reviews. *Human Behavior and Emerging Technologies*, 2024(1):6738001.

Hudon, A. a Stip, E. (2025). Delusional experiences emerging from ai chatbot interactions or "ai psychosis". *JMIR Mental Health*, 12(1):e85799.

Jobin, A., Ienca, M. a Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Picard, R. W., Huffman, J. C. a Nock, M. K. (2018). Digital phenotyping of suicidal thoughts. *Depression and Anxiety*, 35(7):601–608.

Koike, M. a Loughnan, S. (2021). Virtual relationships: Anthropomorphism in the digital age. *Social and Personality Psychology Compass*, 15(6):e12603.

Koike, M., Loughnan, S. a Stanton, S. C. E. (2023). Virtually in love: The role of anthropomorphism in virtual romantic relationships. *British Journal of Social Psychology*, 62(1):600–616.

Loveys, K., Sebaratnam, G., Sagar, M. a Broadbent, E. (2020). The effect of design features on relationship quality with embodied conversational agents: a systematic review. *International Journal of Social Robotics*, 12(6):1293–1312.

Lucas, G. M., Gratch, J., King, A. a Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J. a Morency, L.-P. (2017). Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B. a Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

McRorie, M., Sneddon, I., McKeown, G., Bevacqua, E., De Sevin, E. a Pelachaud, C. (2011). Evaluation of four designed virtual agent personalities. *IEEE Transactions on Affective Computing*, 3(3):311–322.

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M. a Tielman, M. L. (2024). A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4):1–45.

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F. a Twumasi, R. (2025). Delusions by design? how everyday ais might be fuelling psychosis (and what can be done about it).

Nass, C. I. a Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge, MA.

Noor, N., Rao Hill, S. a Troshani, I. (2022). Artificial intelligence service agents: role of parasocial relationship. *Journal of Computer Information Systems*, 62(5):1009–1023.

Norman, D. (1988). *The Design of Everyday Things*. Basic Books.

Ortony, A. (2003). On making believable emotional agents believable. V *Emotions in Humans and Artifacts*, str. 189–211.

Papneja, H. a Yadav, N. (2025). Self-disclosure to conversational ai: a literature review, emergent framework, and directions for future research. *Personal and Ubiquitous Computing*, 29(2):119–151.

Pradhan, A. a Lazar, A. (2021). Hey google, do you have a personality? designing personality and personas for conversational agents. V *Proceedings of the 3rd Conference on Conversational User Interfaces*, str. 1–4.

Rheu, M., Shin, J.-E. Y., Peng, W. a Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction*, 37(1):81–96.

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504.

Shu, C., Lai, K. a He, L. (2026). Human–ai attachment: how humans develop intimate relationships with ai. *Frontiers in Psychology*, 17:1723503.

Simon, H. A. (1968). *The Sciences of the Artificial*. MIT Press.

Skjuve, M., Følstad, A., Fostervold, K. I. a Brandtzæg, P. B. (2021). My chatbot companion – a study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601.

Smestad, T.-M. L. a Volden, F. (2018). Chatbot personalities matters: improving the user experience of chatbot interfaces. V *International Conference on Internet Science*, str. 170–181, Cham. Springer International Publishing.

Tomasello, M. (2010). *Origins of Human Communication*. MIT Press.

Van Pinxteren, M. M., Pluymaekers, M. a Lemmink, J. G. A. M. (2020). Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management*, 31(2):203–225.

Völkel, S. T., Schödel, R., Buschek, D., Stachl, C., Winterhalter, S., Bühner, M. a Hussmann, H. (2020). Developing a personality model for speech-based conversational agents using the psycholexical approach. V *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, str. 1–14.

Wardhana, A. K., Ferdiana, R. a Hidayah, I. (2021). Empathetic chatbot enhancement and development: A literature review. V *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, str. 1–6. IEEE.

Xu, W., Dainoff, M. J., Ge, L. a Gao, Z. (2023). Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai. *International Journal of Human–Computer Interaction*, 39(3):494–518.

Youn, S. a Jin, S.-A. A. (2021). In ai we trust? the effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy". *Computers in Human Behavior*, 119:106721.

Predicting Item Difficulty in German Language Tests: A Machine Learning Approach with Linguistic Features

Ivana Kapounová and Jana Dlouhá

Department of Psychology, Faculty of Arts, Charles University
Celetná 20, 116 42 Praha 1
ivana.kapounova@ff.cuni.cz

Abstract

Item difficulty is central to test validity and fair assessment, yet it is difficult to estimate without large-scale pretesting. Machine learning offers a promising alternative by enabling prediction of item difficulty directly from the text prior to administration. This study replicates and extends previous research on predicting item difficulty in language comprehension tests using machine learning algorithms and linguistic features. Specifically, a modelling approach previously developed for English *maturita* (the standardised school-leaving examination in the Czech Republic) items was applied to German language test items. Item difficulty parameters were estimated using Item Response Theory, and 696 features covering lexical, frequency-based, surface similarity, and readability properties were extracted from the texts. Regularized linear regression models were trained and evaluated using nested cross-validation. Similar to results obtained in the English dataset, the best-performing model (LASSO) achieved a mean RMSE close to the standard deviation of item difficulty and therefore comparable to predicting the mean. This lack of predictive power was further compounded by marked instability across cross-validation folds. Directly applying coefficients trained on English items resulted in substantially larger prediction errors. The results are attributed primarily to the small number of items relative to the high-dimensional feature space. These findings suggest that feature-based machine learning approaches offer limited added value in such settings and are better treated as a complement to, rather than a replacement for, traditional difficulty estimation methods.

1 Introduction

Accurate estimation of item difficulty is central to educational measurement, supporting test validity, score interpretation, and comparability across test forms. It is also essential in modern applications such as computerized adaptive testing, where items must be selected dynamically to match examinee ability and reduce test length. Scalable estimation further enables the construction of balanced tests, detection of potential biases, and more informed decision-making about examinees.

1.1 Traditional Difficulty Estimation

Traditionally, difficulty is established via pretesting and domain expert judgement. Although pretesting yields statistically robust estimates (Fadillah et al., 2023), it is costly, time-consuming, and delays item deployment. Expert judgement, while useful in early stages, is subjective and potentially inconsistent across raters (Wonde et al., 2024).

Yet item difficulty is not arbitrary. It is partly encoded in an item's structural and linguistic properties. Advances in machine learning and natural language processing now make it possible to infer difficulty directly from the text. This shift opens the possibility of predicting item difficulty prior to administration in a scalable and systematic manner, complementing traditional approaches.

1.2 Evolution of Automated Prediction

Early text-based prediction approaches relied on manually engineered surface-level and linguistic features (Loukina et al., 2016), such as sentence length, vocabulary rarity, and syntactic complexity. These features were then used as inputs to classical machine learning models (e.g., linear regression, random forests).

Later developments introduced static word embeddings (e.g., *Word2Vec*, *GloVe*) to capture semantic similarity (Hsu et al., 2018), followed by deep learning architectures such as convolutional neural networks and long short-term memory networks (He et al., 2021; Huang et al., 2017; Xue et al., 2020), which enabled more effective modelling of contextual relationships in text.

1.3 Recent Approaches

Recent advances have shifted the field decisively toward transformer-based architectures and large language models (LLMs), giving rise to three principal paradigms.

First, the most direct approach is to prompt an LLM to predict item difficulty from text (Gombert et al., 2024; Li, Jiao, et al., 2025). While intuitive, this approach typically achieves only moderate accuracy compared to more structured methods (Razavi & Powers, 2025).

Second, LLMs are used as feature extractors, generating cognitively and linguistically informed variables (e.g., reasoning steps or cognitive load) for downstream models (Razavi & Powers, 2025). This approach has been shown to outperform both direct LLM predictions and traditional text-based features (Mojoyinola et al., 2025).

Third, LLMs are used to simulate student responses across proficiency levels (Maeda, 2025; Scarlatos et al., 2025). Item difficulty is subsequently estimated via IRT models fitted to the synthetic data. A key limitation is that LLMs struggle to reflect human capability constraints, and highly capable models tend to overproduce correct responses even when conditioned on lower proficiency (Acquaye et al., 2026; Li, Chen, et al., 2025; Liu et al., 2025).

Beyond these paradigms, LLMs are utilized for data augmentation to enrich training sets with synthetic items and rationales (Feng et al., 2025; Li, Jiao, et al., 2025), as well as for uncertainty estimation, where the model's internal confidence serves as a proxy for predicting human task difficulty (Zotos et al., 2025). A comprehensive overview of these developments is provided by Peters et al. (2025).

1.4 Limitations and Open Challenges

Despite rapid progress in advanced architectures, several important challenges remain unresolved. Among these, interpretability and cross-linguistic generalizability are particularly relevant to this study.

Automated item difficulty prediction involves a persistent trade-off between predictive accuracy and interpretability. Recent LLM-based approaches often achieve superior performance but rely on high-dimensional, opaque representations, limiting transparency. Although post hoc explainable AI methods have been proposed, they offer only partial insight into model decision-making (Dunn et al., 2026). In high-stakes assessment, where transparency is vital for validity and fairness, traditional feature-based models may be therefore preferable. Defined linguistic features allow these approaches to pinpoint difficulty, offering item writers actionable guidance for precise calibration. They are also more robust in small-sample settings, which are typical in high-stakes testing (Peters et al., 2025).

A second open question concerns the cross-linguistic transferability of predictive models. Current evidence in this area is predominantly indirect. For example, recent research suggests that training models jointly on multiple languages can substantially improve cross-lingual performance as they learn global difficulty patterns (Das et al., 2024; Skidmore et al., 2025). For feature-based machine learning models, however, studies explicitly examining transferability are very rare.

1.5 Research Objectives

Two prior studies established a framework for predicting item difficulty from linguistic features using machine learning (Dlouhá, 2025; Štěpánek et al., 2023) but were limited to English *maturita* (the Czech national school-leaving examination) items. This study aims to test the framework's transferability from English to German by replicating it on German *maturita* reading items and to compare these results with the original findings.

2 Methods

2.1 Data

Data were obtained from *CERMAT* German *maturita* tests and response data (spring terms 2016–2024, autumn terms were excluded due to different examinee populations). Only reading-comprehension sections (Parts 5 and 7) were analysed, yielding 45 items per part. A German frequency resource *DeReWo-2014-II* (Institut für Deutsche Sprache, 2014) was used for lexical frequency features.

2.2 Procedure

All analyses were conducted in R (R Core Team, 2023). The replication part followed a three-stage pipeline:

Item difficulty was first estimated via Rasch models (package *eRm*; Mair & Hatzinger, 2007). Responses were analysed per test form without linking.

Items were then segmented into their components (passage, question, key, distractors) from PDFs. A total of 696 base features (including lexical, frequency-based, surface similarity and readability indices) were computed across these components using the *quanteda* (Benoit et al., 2018) and *udpipe* (Wijffels, 2023) packages. Feature–difficulty correlations were modest ($|r| \leq .40$), and substantial multicollinearity combined with a small sample size motivated the use of regularization.

Separate models were subsequently trained for Parts 5 and 7 using ridge, LASSO, and elastic net regression. For that purpose, packages *glmnet* (Friedman et al., 2010) and *caret* (Kuhn, 2008) were used. Nested cross-validation (9 inner and 9 outer folds) served both for unbiased performance estimation and hyperparameter tuning. Before training the model, within each inner loop, the top 100 predictors (by correlation with difficulty) were selected to reduce dimensionality. Performance was assessed via mean RMSE and R^2 from outer folds.

Key differences from prior work include the use of Rasch models (instead of 2PL), a reduced feature set (without *CEFR* or *Word2Vec*), and frequency-based lexical profiling using *DeReWo*.

3 Results

Given the small number of items (n = 45 per part), all results should be interpreted cautiously. In the nested 9×9 cross-validation, each test fold contained only five items, resulting in high variance and limited reliability. Accordingly, the findings are best interpreted as indicative patterns rather than stable effects.

3.1 Model Performance

Across regularization methods (ridge, LASSO, elastic net), differences in predictive performance were marginal, however, LASSO was selected as the primary model due to its feature selection properties and interpretability. Performance metrics are reported in Table 1, which presents mean values and standard deviations across nine outer folds of nested cross-validation. In both test parts, RMSE values were close to the empirical standard deviation of item difficulty (SD_b = 0.67 for Part 5, SD_b = 0.85 for Part 7). This indicates that the models offer only limited improvement over a simple mean-based prediction.

Table 1: Performance Metrics for LASSO Models

	RMSE		R^2	
	Mean	**SD**	**Mean**	**SD**
Part 5	0.69	0.20	.24	.32
Part 7	0.81	0.33	.11	.17

This performance appears slightly weaker than reported by Dlouhá (2025) and broadly comparable to Štěpánek et al. (2023), although differences in feature sets and difficulty characteristics constrain direct comparison.

3.2 English-to-German Transfer

Using the optimal hyperparameters identified in nested cross-validation ($\alpha = 1$, $\lambda = 0.071$), a final LASSO model was trained to examine selected predictors. There was no direct overlap between the predictors retained in this model and those reported by Dlouhá (2025). But given the high multicollinearity among features and the small sample size, this is not unexpected. Under such conditions, LASSO tends to select arbitrary representatives from correlated predictor clusters, limiting interpretability at the level of individual variables.

To examine cross-linguistic transfer at the level of feature space, a model for Part 5 was trained using only the predictors reported by Dlouhá (2025), which resulted in performance comparable to the full feature set (see Table 2 for a comparison of models).

In contrast, transfer at the parameter level was clearly unsuccessful, as applying the original English model with fixed coefficients led to a substantial drop in performance.

Table 2: Comparison of English-to-German Transfer Models (Part 5)

	RMSE	
	Mean	**SD**
Full Feature Set	0.69	0.20
English Features	0.71	0.20
English Coefficients	1.70	-
Baseline	0.67	-

Note. RMSE values for cross-validated models represent mean and standard deviation across 9 folds. The model with English coefficients was evaluated on the full dataset. Baseline corresponds to the standard deviation of item difficulty.

3.3 Feature Selection Stability

Feature selection exhibits limited stability. Approximately 50 predictors consistently ranked among the 100 most highly correlated features in at least 8 of 9 folds, which indicates a partially consistent core set. However, the total number of unique features that met the selection threshold at least once is large (223 for Part 5; 239 for Part 7), reflecting substantial variability, which was more pronounced for Part 7. Given this instability, interpretation of specific predictors within individual models is not warranted.

No single feature shows consistently strong association with item difficulty across both parts. This

absence of robust cross-part predictors is consistent with the low correlation between examinee performance in Parts 5 and 7 (r = .15), suggesting that the two sections might be driven by partially distinct cognitive and linguistic demands. If this is the case, it would mean that item difficulty in this context is not a unitary construct and separate theoretical frameworks and feature engineering strategies may be needed for each.

4 Discussion

This study aimed to test whether a machine learning framework for predicting item difficulty from linguistic features, previously validated on English *maturita* exams by Štěpánek et al. (2023) and Dlouhá (2025), generalises to German, thereby addressing a gap in evidence regarding English-to-German transfer within feature-based difficulty prediction.

The results indicate that linguistic features provide a modest but unreliable predictive signal. Although LASSO performed best among the regularization methods, it showed little improvement over a naïve mean predictor. This pattern mirrors findings from Štěpánek et al. (2023) on English data and suggests that the observed performance ceiling is unlikely to be specific to German items. Similar results have been reported in other small-scale educational datasets, where even more sophisticated models fail to outperform simple baselines (Gombert et al., 2024; Kapoor et al., 2025). Overall, these findings point to limitations in the available data and signal strength rather than in the choice of modelling approach.

In this study, the problem is further amplified by the high dimensionality of the feature space relative to the number of observations. Even with regularisation, the model is prone to overfitting and unstable coefficient estimates, a phenomenon commonly referred to as the curse of dimensionality (Fulari & Rusert, 2024; Peters et al., 2025). Another contributing factor may be the homogeneous design of the *maturita* exam, where the reading items are tailored to a specific proficiency threshold, which constrains variance in both item difficulty and linguistic properties, leaving less signal for models to capture.

Regarding cross-linguistic transfer, the evidence is mixed. At the feature level, a model restricted to predictors identified in prior English-based research (Dlouhá, 2025) achieved performance comparable to the model trained on the full set of available features. However, since both models operate close to the baseline defined by the variance of item difficulty, it likely reflects limited predictive signal rather than successful transfer. In other words, two uninformative models will always converge on approximately the same RMSE regardless of their feature content. The feature-level transfer question therefore remains open rather than confirmed or disconfirmed. At the parameter level, directly applying the English-trained coefficients to German data produced a substantially higher RMSE confirming that parameter estimates are language-specific and retraining is necessary.

Three additional limitations constrain the validity of the results. First, not all features from Dlouhá (2025) were extracted (e.g., CEFR lists, Dale-Chall index, and Word2Vec semantic similarity), which may further reduce comparability with prior work. Second, multicollinearity complicates the interpretation of selected predictors, as several features capture overlapping information (e.g., features reflecting item length, such as raw counts). Third, the cross-linguistic scope of this study is restricted to a comparison of English and German languages and whether the observed performance patterns extend to other languages remains an open empirical question.

To address these limitations, future research should rely on substantially larger datasets when applying this methodology. In contexts where larger samples are not available, the number of predictors should be reduced by prioritising theoretically grounded features, or by applying more advanced feature engineering and dimensionality reduction techniques. Predictive performance may also benefit from incorporating more recent interpretable approaches, such as features derived from LLMs (e.g., indicators of cognitive load or required reasoning complexity). Finally, for cross-linguistic comparisons, it may be useful to adopt approaches based on explicitly formulated and theoretically informed hypotheses, as this could lead to more interpretable and convincing insights into language-specific patterns.

5 Conclusion

Recent advances in machine learning, particularly in large language models, have made automated item difficulty prediction a rapidly developing area. This study aimed to replicate and extend prior research on English *maturita* items (Dlouhá, 2025; Štěpánek et al., 2023) by evaluating the same framework on German data and examining its transferability between these two languages.

The results show that predictive performance remains limited across all conditions, and instability in feature selection prevents meaningful interpretation of individual predictors. These findings are consistent with earlier work and suggest that the main constraint lies in small item sample size. In such settings, these approaches may offer only limited added value and are better viewed as a complement to traditional difficulty estimation methods.

References

Acquaye, C., Huang, Y. T., Carpuat, M., & Rudinger, R. (2026). *Take out your calculators: Estimating the real difficulty of question items with LLM student simulations*. arXiv. https://doi.org/10.48550/arXiv.2601.09953

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774–774. https://doi.org/10.21105/joss.00774

Das, R., Hristov, S., Li, H., Dimitrov, D., Koychev, I., & Nakov, P. (2024). EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7768–7791). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.420

Dlouhá, J. (2025). *Item response theory and machine learning in modeling difficulty of educational assessments* [Doctoral dissertation]. Univerzita Karlova, Filozofická fakulta.

Dunn, K. J., Skidmore, L., & Rogers, T. (2026). When measurement meets machine learning: Interpretability and scalability in modelling item difficulty for language assessment. *Frontiers in Education*, *11, 1740237*. https://doi.org/10.3389/feduc.2026.1740237

Fadillah, S. M., Ha, M., Nuraeni, E., & Indriyanti, N. Y. (2023). Exploring confidence accuracy and item difficulty in changing multiple-choice answers of scientific reasoning test. *Malaysian Journal of Learning and Instruction (MJLI)*, *20*(2), 319–341. https://doi.org/10.32890/mjli2023.20.2.5

Feng, W., Tran, P., Sireci, S., & Lan, A. S. (2025). Reasoning and sampling-augmented MCQ difficulty prediction via LLMs. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds), *Artificial Intelligence in Education* (pp. 31–45). Springer Nature Switzerland.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Fulari, R., & Rusert, J. (2024). Utilizing machine learning to predict question difficulty and response time for enhanced test construction. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 528–533). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.45/

Gombert, S., Menzel, L., Di Mitri, D., & Drachsler, H. (2024). Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 483–492). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.40/

He, J., Peng, L., Sun, B., Yu, L., & Zhang, Y. (2021). Automatically predict question difficulty for reading comprehension exercises. *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 1398–1402. https://doi.org/10.1109/ICTAI52525.2021.00222

Hsu, F.-Y., Lee, H.-M., Chang, T.-H., & Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, *54*(6), 969–984. https://doi.org/10.1016/j.ipm.2018.06.007

Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., & Hu, G. (2017). Question difficulty prediction for reading problems in standard tests. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10740

Institut für Deutsche Sprache. (2014). *DeReWo-2014-II-Hauptarchiv-STT.100000: Korpusbasierte Wortformenliste DEREWO, v-xxx, mit Benutzerdokumentation* [Data set]. Institut für Deutsche Sprache. https://www.ids-mannheim.de/kl/derewo/

Kapoor, R., Truong, S. T., Haber, N., Ruiz-Primo, M. A., & Domingue, B. W. (2025). *Prediction of item difficulty for reading comprehension items by creation of annotated item repository*. arXiv. https://doi.org/10.48550/arXiv.2502.20663

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Li, M., Chen, H., Xiao, Y., Chen, J., Jiao, H., & Zhou, T. (2025). *Can LLMs estimate student struggles? Human-AI difficulty alignment with proficiency simulation for item difficulty prediction*. arXiv. https://doi.org/10.48550/arXiv.2512.18880

Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., & Lissitz, R. W. (2025). Item difficulty modeling using fine-tuned small and large language models. *Educational and Psychological Measurement*, *85*(6), 1065–1090. https://doi.org/10.1177/00131644251344973

Liu, Y., Bhandari, S., & Pardos, Z. A. (2025). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, *56*(3), 1028–1052. https://doi.org/10.1111/bjet.13570

Loukina, A., Yoon, S.-Y., Sakano, J., Wei, Y., & Sheehan, K. (2016). Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In Y. Matsumoto & R. Prasad (Eds), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3245–3253). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1306/

Maeda, H. (2025). Field-testing multiple-choice questions with AI examinees: English grammar items. *Educational and Psychological Measurement*, *85*(2), 221–244. https://doi.org/10.1177/00131644241281053

Mair, P., & Hatzinger, R. (2007). Extended rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Mojoyinola, M., Kehinde, O. J., & Tang, J. (2025). Enhancing item difficulty prediction in large-scale assessment with large language models. In J. Wilson, C. Ormerod, & M. Beiting Parrish (Eds), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress* (pp. 218–222). National Council on Measurement in Education (NCME). https://aclanthology.org/2025.aimecon-wip.27/

Peters, S., Zhang, N., Jiao, H., Li, M., & Zhou, T. (2025). Review of text-based approaches to item difficulty modeling in large-scale assessments. In J. Wilson, C. Ormerod, & M. Beiting Parrish (Eds), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers* (pp. 37–47). National Council on Measurement in Education (NCME). https://aclanthology.org/2025.aimecon-sessions.4/

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.0) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Razavi, P., & Powers, S. J. (2025). *Estimating item difficulty using large language models and tree-based machine learning algorithms*. arXiv. https://doi.org/10.48550/arXiv.2504.08804

Scarlatos, A., Fernandez, N., Ormerod, C., Lottridge, S., & Lan, A. (2025). SMART: Simulated students aligned with item response theory for question difficulty prediction. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 25071–25094). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1274

Skidmore, L., Felice, M., & Dunn, K. (2025). Transformer architectures for vocabulary test item difficulty prediction. In E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds), *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)* (pp. 160–174). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.bea-1.12

Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, *11*(19), 4104. https://doi.org/10.3390/math11194104

Wijffels, J. (2023). *udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'UDPipe' 'NLP' toolkit* (Version 0.8.11) [Computer software]. https://CRAN.R-project.org/package=udpipe

Wonde, S. G., Tadesse, T., Moges, B., & Schauber, S. K. (2024). Experts' prediction of item difficulty of multiple-choice questions in the Ethiopian Undergraduate Medicine Licensure Examination. *BMC Medical Education*, *24*(1), 1016. https://doi.org/10.1186/s12909-024-06012-x

Xue, K., Yaneva, V., Runyon, C., & Baldwin, P. (2020). Predicting the difficulty and response time of multiple choice questions using transfer learning. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, & T. Zesch (Eds), *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 193–197). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.20

Zotos, L., Rijn, H. van, & Nissim, M. (2025). Are you doubtful? Oh, it might be difficult then! Exploring the use of model uncertainty for question difficulty estimation. In C. Mills, G. Alexandron, D. Taibi, G. L. Bosco, & L. Paquette (Eds), *Proceedings of the 18th International Conference on Educational Data Mining* (pp. 77–89). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.15870153

Explorations in Predictive Coding Models of Hallucinations

Milica Kiš, Martin Takáč

FMFI, UNIBA, Bratislava
Fakulta matematiky, fyziky a informatiky
Univerzity Komenského
Mlynská dolina F1
842 48 Bratislava
Email: milica.kis@fmph.uniba.sk

Abstract

In this paper, we explore the emergence of visual hallucinatory-like structures in a hierarchical predictive coding model trained on the MNIST digit dataset. Adding precision to the model enables exploration of the role of priors in hallucinations. Forcing sensory precision to low values corresponds to regarding the sensory data as unreliable. Conversely, precision in the upper layers driven to disproportionately high values yields strong priors, corresponding to rigid beliefs characteristic of psychiatric disorders. The aberrant precision weighting is taken as the mechanism underlying the hallucinations. Hallucinatory-like behavior can be induced either by clamping the top layer to a specific digit or by letting the system evolve its internal landscape from an unconstrained internal state. The latter would correspond to spontaneous hallucinations in the absence of sensory input, while the former would result in the generative model maintaining a strong belief despite the sensory evidence.

1 Introduction

Recently, predictive coding has shown promise as an explanatory and exploratory framework for phenomena in computational psychiatry. This influential theory in neuroscience (Friston and Kiebel, 2009) has been used to explain various cognitive phenomena, including those in the domain of human vision.

In predictive coding, processing of neural signals approximates Bayesian inference on the latent causes of sensory data. Observations (likelihoods) are weighed against expectations (priors) to arrive at better beliefs (posterior probabilities or posteriors). Prediction errors signal the discrepancy between likelihoods and priors, which guides adaptive belief updating. The goal of this process is to minimize the free energy. The theory assumes a hierarchical brain structure and bidirectional transmission of neural messages (bottom-up and top-down). From the standpoint of computational psychiatry, positive psychotic symptoms, such as hallucinations and delusions, may be seen as profound disruptions in forming, maintaining, and updating our beliefs about the world.

In this paper, we focus on visual hallucinatory-like states of indeterminate etiology.[1] In a general sense, hallucinations signify percepts arising without the corresponding external stimulus. The other most common type is auditory (verbal) hallucinations (AVH), i.e., hearing voices (Corlett et al., 2019).

In the current literature, there are two opposing hypotheses about the nature of priors associated with hallucinations (Corlett et al., 2019; Sterzer et al., 2018). The strong prior hypothesis states that hallucinations result from enhanced top-down predictive signaling (i.e., increased prior precision) relative to sensory-level activity. Therefore, perception relies more on beliefs than on the sensory input. Conversely, a weak prior account (i.e., lower precision) posits relying more on the ambiguous or noisy input from the sensory level, exerting a disproportionate influence on perception. However, these two hypotheses are not mutually exclusive and may operate simultaneously under different conditions.

The additional complexity is that different priors may be associated with different sensory modalities and hierarchical levels of brain organization. Higher-level associative cortical regions exhibit a higher density of recurrent connections than lower sensory regions, with implications for excitation-inhibition dynamics (Corlett et al., 2019). Consequently, priors that are active at different levels have implications for precision dynamics.

2 Hierarchical predictive coding (HPC) network

To test specific hypotheses about the emergence of hallucinatory-like behavior in the visual system arising from aberrant precision weighting, we introduce a hierarchical predictive coding (HPC) architecture based on the model by Tscshantz et al. (2023), with modifications

[1] Apart from psychiatric disorders, such as schizophrenia and post-traumatic stress disorder, hallucinations may be associated with neurological conditions, such as Alzheimer's disease or epilepsy, eye conditions, or may even appear or be induced in the healthy population (Corlett et al. (2019).

and simplifications (Fig. **??**). The current model disregards the amortized feedforward route, but adds dynamic precision updating, which allows direct manipulation of precision weighting at different hierarchical levels during hallucination experiments.

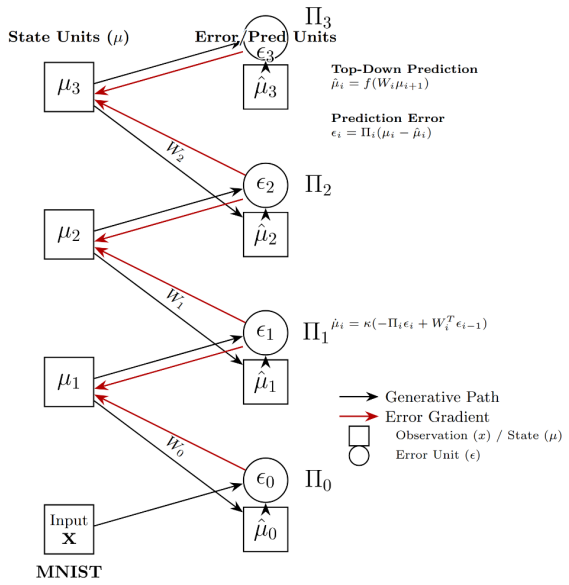


Fig. 1: Hybrid predictive coding architecture with precision matrices. Based on Tscshantz et al. (2023).

2.1 Model description

HPC models perception as an iterative process of Bayesian inference across hierarchical levels.

Generative, top-down pathway. Each hierarchical layer attempts to predict the activity of the layer below through top-down generative connections. The discrepancy between the prediction and the actual activity (or sensory input, in the case of the lowest, sensory layer) is encoded as a prediction error, which is propagated upward to update latent beliefs.

Iterative inference. Inference proceeds iteratively by minimizing variational free energy through updates of latent states. The bidirectional propagation of errors over a number of inference steps enables the network to gradually settle into the final, internally consistent interpretation of the sensory input. The intermediate step (in our model) would be the update of the precisions once the latent states settle, with weights learning from the precision-weighted errors as the final step.

2.1.1 HPC update formulas

Formulas for updates in HPC framework are based on the joint optimization objective of free variational energy minimization. Iterative inference, precision and weight updates are therefore derived by taking the corresponding gradient on variational free energy, with respect to latent states (μ), weights (W) and precision (inverse variance) π, respectively.

2.1.2 Iterative inference

Latent states are updated according to the following formula:

$$\dot{\mu} = -\kappa \frac{\partial \mathcal{F}}{\partial \mu} = -\kappa \left(\pi_p \varepsilon_p - \frac{\partial f_\theta(\mu)}{\partial \mu}^\top \pi_l \varepsilon_l \right) \quad (1)$$

where $\varepsilon_p = \mu - \bar{\mu}$ and $\varepsilon_l = x - f_\theta(\mu)$ and the precisions π_p and π_l correspond to the inverse variances at the two neighbouring layers. The non-linearities are represented with f_θ. The inference rate is denoted by κ.These formulas were adapted from Tscshantz et al. (2023).

2.1.3 Weight updates

Weight updates correspond to local Hebbian learning. The following update formula is obtained by taking the gradient with respect to (W):

$$\Delta W_l = \eta(\pi_l \varepsilon_l)\mu_{l+1}^\top \quad (2)$$

The first term $(\pi_l \varepsilon_l)$ corresponds to the postsynaptic error, i.e., error scaled by reliability ($\varepsilon_l \pi_l$), since $\pi = \Sigma^{-1}$, i.e., the precision matrix is the inverse of a covariance matrix. The error term shows what is surprising, and the precision shows how much we trust that surprise (magnitude). The second term, $\mu_{l+1}^\top$, represents the presynaptic activity of the layer above. Learning rate for weights is denoted by η.

2.2 Network architecture

We implemented a 4-layer hierarchical PC model with interpretable inference dynamics to replicate the normal activity in the visual cortex (with 4 layers loosely interpreted as corresponding to areas V1, V2, V4, and IT).[2] The lowest, sensory input layer consists of 784 neurons (corresponding to 28 x 28 pixels input MNIST digit images). The second and third layers contain 256 and 64 neurons, respectively, while the top (output) layer has 10 neurons, corresponding to the MNIST digit identity (classes). The reduction from 256 to 64 units implements a dimensional bottleneck consistent with increasing abstraction along the ventral visual stream.

[2]This mapping is conceptual and does not reflect strict anatomical correspondence.

2.3 Initialization

Generative (top-down) weights are initialized by sampling from a normal distribution (due to the theoretical Gaussian assumptions). Precisions are initiated as vectors of ones.

Clamping. The highest-level latent representation is clamped to the target label (a one-hot vector) during training, thereby implementing supervised learning.

2.4 Network performance metrics

Under similar (Gaussian) assumptions, Mean Square Error (MSE, summed across all layers) serves as a proxy for minimizing the variational free energy across epochs. We termed this approximation the free energy drop. Reconstruction quality is estimated from the lowest, sensory layer. It is presented as the percentage of drop in MSE relative to the first epoch (assuming that the energy is minimized). The drop in free energy and MSE during regular inference is 95.79%, and 98.31%, respectively. Reconstructions allow visual comparison with the original MNIST digits (Fig. 2).

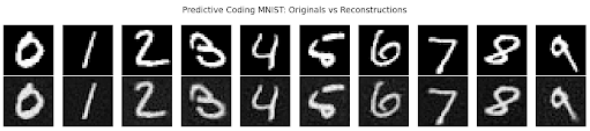


Fig. 2: Original MNIST digits vs. reconstructions (generative model).

2.4.1 Hyperparameters

The following hyperparameters were used: (1) **inference rate** (κ), controlling the step size of latent state updates, a fast component (set to 0.001), (2) **learning rate for weights** (η), set to a small value, i.e., $\eta = 0.001$, a slow learning component, (3) **number of inference steps** (set to 30), (4) **number of training epochs** (set to 100), (5) **learning rate for precisions** (K) (set to 0.005).

2.4.2 Precisions in PC

In predictive coding, precisions (π) reflect the brain's confidence in the prediction errors at each hierarchical level. A high precision means the system strongly trusts the error signal from that layer, and thus assigns more weight to that error when updating its beliefs (latent states). Conversely, low precision would decrease the weighting of the prediction error. Precision is formalized as the inverse of variance or negative entropy (Corlett et al., 2019). Changes in precision are implemented biologically via neuromodulators like dopamine, noradrenaline, or acetylcholine (Feldman and Friston, 2010). Precision is thought to be encoded by the postsynaptic gain of neurons reporting the prediction error. The gain control adjusts the weighting of precision errors. Synaptic gain is a fundamental biophysical property of neuronal function, and it describes the sensitivity of a postsynaptic neuron's membrane potential to variations of strength of the presynaptic input (Nour and Dolan, 2022).

Relationship between precision, priors and posteriors. In HPC, predictions are encoded as predictive signals from the higher levels of hierarchy to the lower, sensory levels. In the event of sensory data contradicting this top-down information, a prediction error signal is sent in the opposite, bottom-up direction towards the higher levels of the hierarchy. Formalized as Bayesian inference, predictions (priors) and sensory data (likelihood) are represented in the form of probability distributions. Posterior belief is shifted toward the prior or sensory evidence (likelihood) in proportion to their relative precision (as the inverse of variance).

Dynamic precision updating. To account for context-dependent uncertainty, we extend the Hybrid Predictive Coding (HyPC) framework proposed by Tscshantz et al. (2023) by introducing dynamic precision updating. Instead of being static, precisions are iteratively optimized via gradient descent on the variational free energy objective function. The precision, being the inverse of the variance $\pi = \frac{1}{\sigma^2}$, can be iteratively updated at each level l using the formula:

$$\Delta\pi_l = -K * \left(\epsilon_l^2 - \frac{1}{\pi_l}\right) , \tag{3}$$

where the squared error signal ϵ_l^2 (the difference between expectation and the bottom-up signal at level l) is conveniently taken to represent the variance. Initialized as identity vectors ($\boldsymbol{\pi}_l = \mathbf{1}$), this formulation acts as a running, sliding-window estimator of error variance. Consequently, precision naturally increases when prediction errors are small (signaling high model confidence) and decreases when errors are large (signaling ambiguity or noise). The order of updates is also important. First, the inference updates settle latent states, then the precisions adapt to current uncertainty, and, finally, the weights learn from precision-weighted errors.

3 Precision experiments

Replicating hallucinatory-like effects by manipulating precisions.

3.1 Manipulating precisions across layers

Adding precision to the model enables exploration of the role of priors in hallucinations. Forcing sensory precision to low values corresponds to regarding the sensory data as unreliable. Conversely, precision in

the upper layers driven to disproportionately high values yields strong priors, corresponding to rigid beliefs characteristic of psychiatric disorders. We created an imbalanced hierarchy as a way to implement aberrant precision weighting, which is taken as the mechanism underlying hallucinations. In modeling terms, emerging hallucinatory-like structures can be induced either by clamping the top layer to a specific digit or by letting the system explore its unconstrained internal landscape. The former would result in the generative model maintaining a strong belief despite the sensory evidence (blank input in this case, see Fig. 3), while the latter would correspond to spontaneous hallucinatory-like activity in the absence of sensory input (see Fig. 4). The rationale behind this choice is that hallucinations (including visual) are usually defined as the false sensory perception in the absence of external sensory stimuli.

This model enables testing different hypotheses within the same architecture (using the same inference dynamics). Prior dynamics can be shifted by assigning different precision values to different layers, sensory input can be either blank, regular, or corrupted or noisy, while clamping enables constraining or freeing the inference dynamics. Therefore, we can operationalize the concepts of strong or weak priors by adjusting the elements listed above.

3.2 Two hallucinatory-like regimes

3.2.1 Top-down hallucinatory-like behavior (strong priors)

In this regime, we are clamping the top layer to a specific digit. The network is generating data from internal strong priors and is managing to produce meaningful reconstructions, despite the blank input.

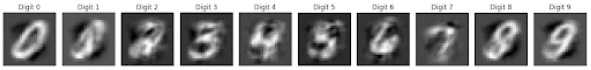


Fig. 3: Hallucinated digits from strong priors

3.2.2 Spontaneous hallucinatory-like behavior

Network samples its own generative model in five iterations in unconstrained generative regime, in the absence of sensory input. The resulting reconstructions indicate that the system is struggling to reconstruct a meaningful output.

Note. We also observed the negative drop in free energy, i.e., the increase, during these experiments. This is expected because the network is generating an image (i.e., a digit) from a blank input, leading to a large prediction error and consequently an increase in free energy.



Fig. 4: Spontaneous hallucinatory-like behavior

4 Logistic regression for estimating the classification confidence

In order not to be relying solely on the observed visual similarity with the original MNIST digits, we used Logistic Regression for estimating the classification confidence (i.e., probability distribution of digits given the hallucinated image). Classification confidence was higher in the case of top-down strong prior condition (with the exception of digits 3, 8, and 9) (Fig. 5), while for spontaneous hallucinatory-like probabilities were more distributed, which might suggest that spontaneous generation is more ambiguous, and that the system is trying to explore various internal landscapes (Fig. 6)). The resulting ambiguity for digits 3, 8, and 9 might lead to speculation about the perceived similarities of rounded and semi-rounded shapes in these figures, which also might have been encoded in their latent representations.

5 Conclusion

Exploring hallucinatory-like behavior in a hierarchical predictive coding model approximating human vision offers a glimpse into internal dynamics of aberrant precision-weighting, which is thought to underlie many psychiatric conditions. Although these deficits may be tested in experimental conditions (mismatch negativity, sensory attenuation), computational modeling offers an alternative method of testing these hypotheses.

Future directions. Another possible direction in our research is testing the hypothesis of weak priors. Under these assumptions, hallucinatory-like behavior arises from insufficient top-down constraints on perception, such that noisy or ambiguous sensory input is not adequately regularized by prior expectations. In the model, we could achieve this by reducing the precisions of higher hierarchical layers (which would correspond to a weak prior) and providing a noisy or corrupted input, allowing sensory ambiguities to drive inference.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D. and Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in psychiatry*, 4:47.

Bogacz, R. (2017). A tutorial on the free-energy frame-

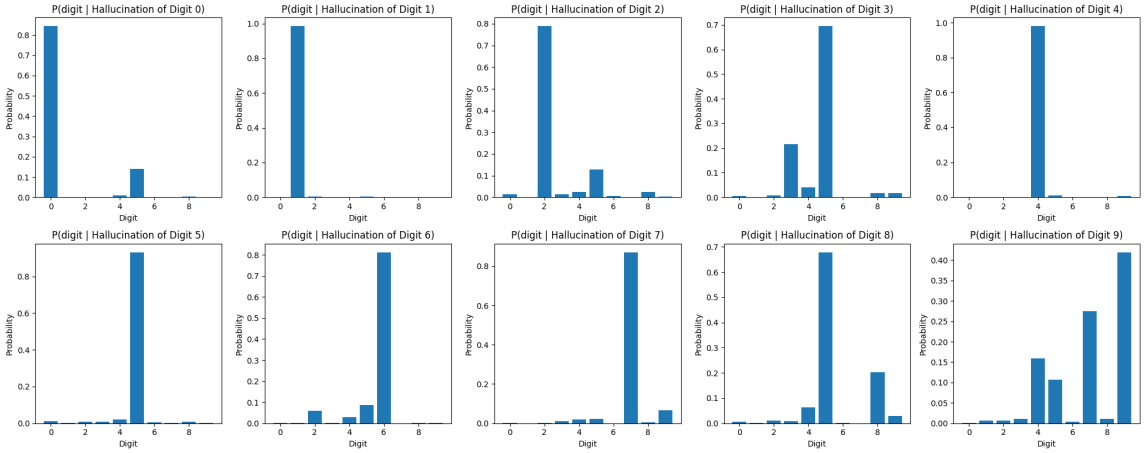


Fig. 5: Probability distributions in strong prior top-down condition

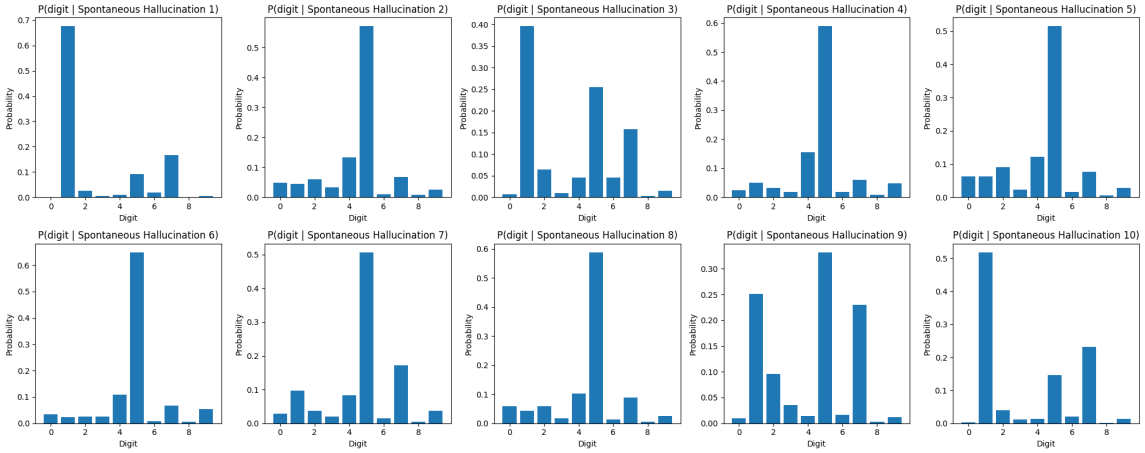


Fig. 6: Probability distributions in spontaneous condition

work for modelling perception and learning. *Journal of mathematical psychology*, 76:198–211.

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K. and Powers, A. R. (2019). Hallucinations and strong priors. *Trends in cognitive sciences*, 23(2):114–127.

Feldman, H. and Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 4:215.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221.

Mancini, V. and Nour, M. M. (2025). If mismatch negativity is the answer, what is the question? on the nature of predictive coding abnormalities in psychosis. *Biological Psychiatry Global Open Science*, 5(1):100412.

Nour, M. M. and Dolan, R. J. (2022). Synaptic gain abnormalities in schizophrenia and the potential relevance for cognition. *Biological psychiatry*, 91(2):167–169.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M. and Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological psychiatry*, 84(9):634–643.

Tscshantz, A., Millidge, B., Seth, A. K. and Buckley, C. L. (2023). Hybrid predictive coding: Inferring, fast and slow. *PLoS computational biology*, 19(8):e1011280.

Izhikevich Model of a Spiking Neuron: A Qualitative Analysis Through the Phase Plane

Hana Kollárová, BSc.

Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics, Comenius University of Bratislava
KAI FMFI UK, Mlynska dolina, 842 48 Bratislava
kollarova223@uniba.sk

Abstract

This paper discusses the biological basis for computational models and introduces the Izhikevich model. The Izhikevich model is a simplified spiking neuron model that combines biological realism with computational efficiency. While inspired by the dynamical behaviour captured by the Hodgkin–Huxley framework, it reduces the description to two coupled differential equations and a reset mechanism. Parametric sensitivity of the Izhikevich model will be explored through input current alterations. Specifically, experiments were conducted using constant external current, single-pulse stimulation with varying onset time, single-pulse stimulation with varying amplitude, and double-pulse stimulation with varying inter-pulse delay. The resulting voltage traces and phase-plane trajectories were analysed to characterise excitability thresholds, tonic firing, state-dependent spike generation, and refractory dynamics. The results qualitatively correlate with the behaviour of biological neurons, proving that the Izhikevich model is able to replicate their characteristics appropriately.

1 Introduction

Computational neuroscience uses mathematical models to study how neurons generate and transmit electrical activity. Because the underlying biophysical mechanisms involve many interacting processes, models are typically simplified, with the level of abstraction chosen based on the problem being studied (Cessac and Samuelides, 2007). Detailed conductance-based models can reproduce ionic dynamics with high accuracy, but their complexity limits their use in large-scale simulations and systematic analysis. For this reason, reduced models are commonly used when the focus is on qualitative behaviour such as spike generation and pattern formation.

A widely used example is the Izhikevich neuron model, introduced by Eugene M. Izhikevich (Izhikevich, 2003), which combines low computational cost with the ability to reproduce multiple firing patterns observed in biological neurons. The model is defined as a two-dimensional system:

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I, \tag{1}$$

$$\frac{du}{dt} = a(bv - u), \tag{2}$$

with the after-spike reset:

$$\text{if } v \geq 30 \text{ mV, then} \quad \begin{cases} v \leftarrow c, \\ u \leftarrow u + d. \end{cases} \tag{3}$$

In this formulation, v represents the membrane potential and u is a recovery variable capturing slower processes such as ionic currents and membrane adaptation. The external input I drives the system, while the parameters a, b, c, and d determine the qualitative firing regime. Parameter a controls the time scale of recovery, b determines the coupling between membrane voltage and recovery dynamics, c specifies the membrane potential after a spike, and d governs the post-spike increase of the recovery variable. Unlike more detailed conductance-based approaches, these parameters are not tied to specific ion channels but instead control the system's dynamical response.

Compared to the Hodgkin–Huxley model (Izhikevich, 2007), which explicitly models ion channel kinetics using four coupled non-linear equations, the Izhikevich model reduces the system to two dimensions while retaining key behaviours such as threshold dynamics, spike generation, and refractory effects. This reduction makes it possible to explore how qualitative changes in input affect neuronal activity without the computational overhead of biophysical detail. Reduced models become particularly relevant when extending the analysis beyond single neurons. In this context, the Izhikevich model provides a practical compromise: it is simple enough for large-scale simulations, but still expressive enough to capture distinct firing patterns.

This paper examines how the behaviour of the Izhikevich model changes under different forms of external input. The analysis focuses on how variations in the input current I influence spike generation and overall system dynamics. Specifically, constant current stimulation, single-pulse stimulation, and double-pulse

stimulation are considered, with the goal of relating observed behaviours to the underlying dynamical structure of the model.

2 Method

The parameter set for the model is the following:

$$a = 0.22, \quad b = 0.183, \quad c = -65.81, \quad d = 9.18, \tag{4}$$

chosen for *qualitative coverage* of excitability phenomena (subthreshold transients, single spikes, spike trains, refractory effects, and strong deposition), rather than to reproduce a canonical cell type from Izhikevich (2003). This set extends earlier work done by the authors of this paper on the phase plane analysis of the FitzHugh-Nagumo model (Kollárová, 2025): the FitzHugh-Nagumo Hopf threshold ($I \approx 0.34$) (Musacchio, 2025b) maps to an Izhikevich saddle-node rheobase near $I = 4$ for canonical RS parameters; here $b = 0.183$ shifts the analytic rheobase to $I_{\text{rheo}} = (5 - b)^2/0.16 - 140 \approx 5.0$.

The initial conditions for the system are set to:

$$v_0 = -55.108; u_0 = 0.183 \cdot (-55.108) \approx -10.085.$$

For I_{ext} below rheobase, the $\dot{v} = 0$ and $\dot{u} = 0$ nullclines intersect on the lower branch at a stable node. At $I_{\text{ext}} = I_{\text{rheo}}$ these fixed points collide and annihilate in a *saddle-node bifurcation*; above rheobase no fixed point remains and repetitive firing occurs. In the excitable regime just above threshold, trajectories briefly "feel" the disappeared equilibrium (a *ghost equilibrium*), producing slow passage and sharp threshold sensitivity to pulse timing.

For qualitative analysis, the nullclines of the system are derived directly from the parameterised equations. The $\dot{v} = 0$ nullcline is given by:

$$u = 0.04v^2 + 5v + 140 + I_{\text{ext}}, \tag{5}$$

which forms a quadratic curve in the (v, u) phase plane. The $\dot{u} = 0$ nullcline follows as:

$$u = 0.183v, \tag{6}$$

which is linear with slope 0.183.

The intersection of these nullclines determines the fixed points of the system. However, under the chosen parameters, the system does not settle into a stable equilibrium and instead exhibits spike-driven dynamics governed by the reset condition.

There will be three parts to the experiment. In the first part, the external current will be kept constant. In the second part, the external current will be rising linearly, and in the third, it will be limited to a specific time interval $t_0 - t_1$, and, therefore, will be a pulse. There will be either exactly one or two pulses at any given time, and they will be of a set external current. In terms of units, the current is expressed in the same abstract units as in (Izhikevich, 2003). These values are scaled inputs directly added to the voltage dynamics and therefore do not correspond to a physical current unless a specific calibration is introduced.

The experiment setup will be as follows:

(1) Constant current. $I_{\text{ext}} \in (0, 500)$ for $t \in [0, 500]$.

(2) Single pulse, timing dependence. Rectangular pulse $I_{\text{ext}} = 22$ on $t \in [t_0, t_0 + 1]$ for $t_0 = 0, 1, \ldots, 9, 10, 13, 15$ (13 windows), simulation length 250.

(3) Single pulse, amplitude dependence. Fixed window $t \in [13, 14]$; $I_{\text{ext}} \in \{4, 8, 12, 14, 16, 20, 24\}$.

(4) Double pulse, refractory period. Two pulses of amplitude 7 on $[10, 15]$ and a second window separated by gap $\Delta t \in \{0, 5, 10, 25\}$: second pulse on $[15, 20]$, $[20, 25]$, $[25, 30]$, or $[40, 45]$.

3 Results

3.1 Results Overview of the Izhikevich Model

The Izhikevich model was systematically evaluated under constant and pulsed external current inputs in order to assess its ability to reproduce key biological neuronal properties. Across all regimes, the model consistently demonstrated classical excitability, including subthreshold oscillations, spike generation, spike trains, and refractory behaviour.

3.1.1 Constant External Current

For constant external current, the system showed a clear transition between resting-like behaviour and repetitive firing. At low input levels, only small sub-threshold fluctuations around the resting state were observed. These sub-threshold fluctuations happened mostly because of the introduction of the ghost equilibrium, as the Izhikevich model does not have any noise introduced in the system. Therefore, with too low of a current, the system is rather unresponsive.

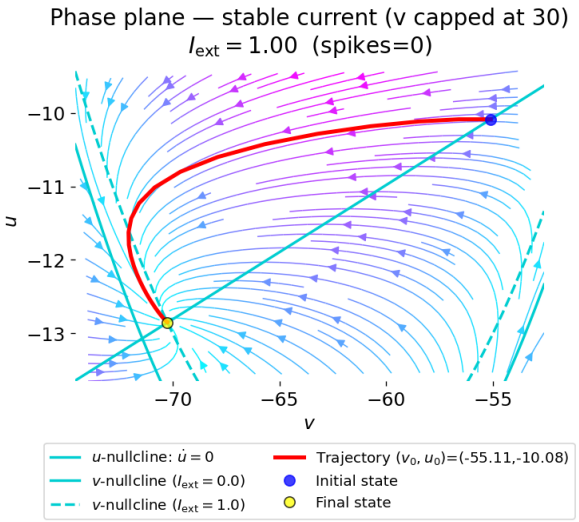


Fig. 1: Stable external current of height $I_{ext} = 1.0$. Only a small oscillation happened, as the external current was too low to produce an action potential.

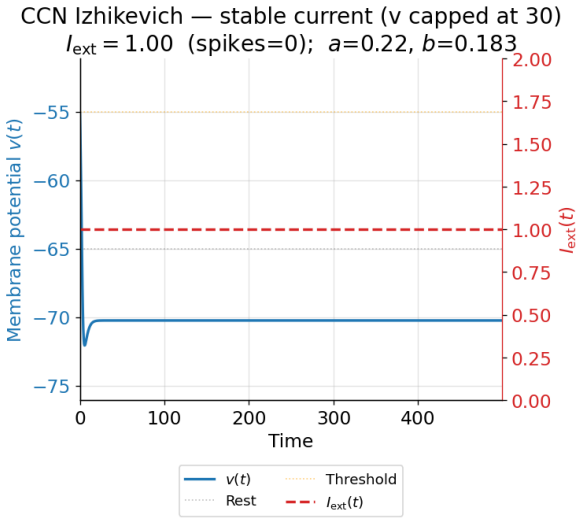


Fig. 2: No action potential occurred. The "ghost equilibrium" that was set up can be seen at time 0, as there is an initial displacement from the actual equilibrium.

With increasing current, the system crossed a firing threshold and entered a stable limit cycle regime. In this regime, sustained spike trains were produced, confirming tonic spiking behaviour. The transition from no firing to repetitive firing demonstrated the presence of a well-defined excitability threshold, consistent with biological neurons.

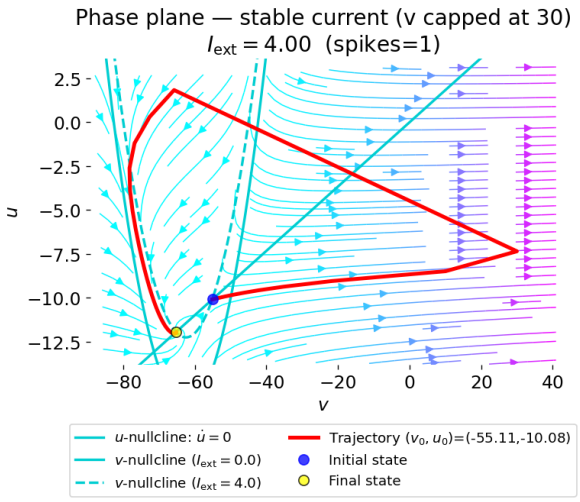


Fig. 3: Stable external current of height $I_{ext} = 4.0$. Exactly one action potential occurred, even though the current was introduced permanently.

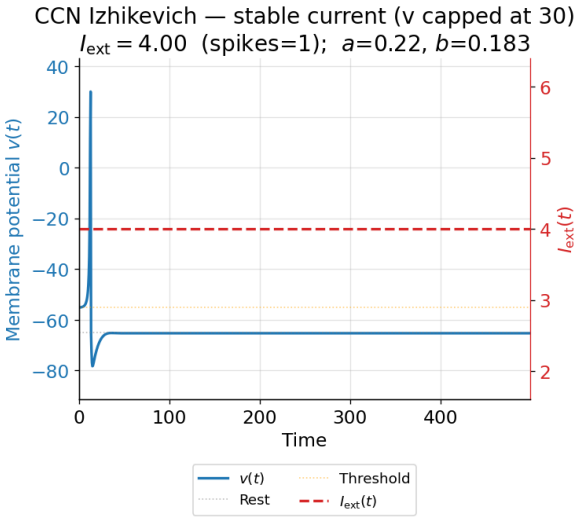


Fig. 4: The resulting singular action potential.

For very high constant input, the system enters an excitation-dominated regime rather than the orderly firing observed at moderate currents (e.g. $I_{\text{ext}} = 8$, where inter-spike intervals allow visible recovery toward the reset level). At $I_{\text{ext}} = 40$ and especially $I_{\text{ext}} = 500$, the membrane potential is driven back toward threshold almost immediately after each reset: spike trains become extremely dense (143 and 500 spikes over the simulation window, respectively), and the trajectory spends little time on the lower branch of the phase plane where refractory dynamics normally develop. Because the reset is instantaneous and independent of how strongly the neuron is depolarised, it overwrites both relative and absolute refractory effects that would otherwise suppress firing; the slow recovery variable u cannot enforce a biologically meaningful refractory interval when I_{ext} is

large enough to re-excite the cell within a single integration step.

This contrasts with biological tissue under extreme stimulation, where sustained depolarisation typically leads to *depolarisation block*, ionic rundown, or irreversible damage rather than an unbounded acceleration of action-potential firing (Gazzaniga et al., 2018). The behaviour is thus useful for illustrating how the Izhikevich reset mechanism supports rich excitability at moderate input, but also exposes a clear limitation: without additional constraints on v, u, or I, the model cannot reproduce the failure modes of real neurons.

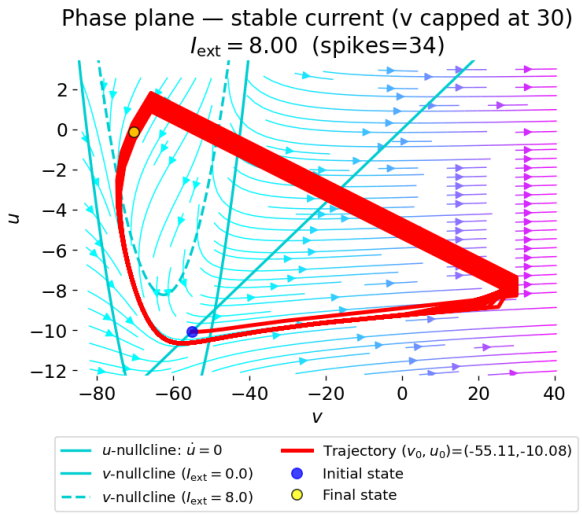


Fig. 5: A spike train occurs when $I_{ext} = 8.0$, which can be seen by repetitive expeditions of the system through the phase plane. 34 spikes occur over a period of 500 units of time.

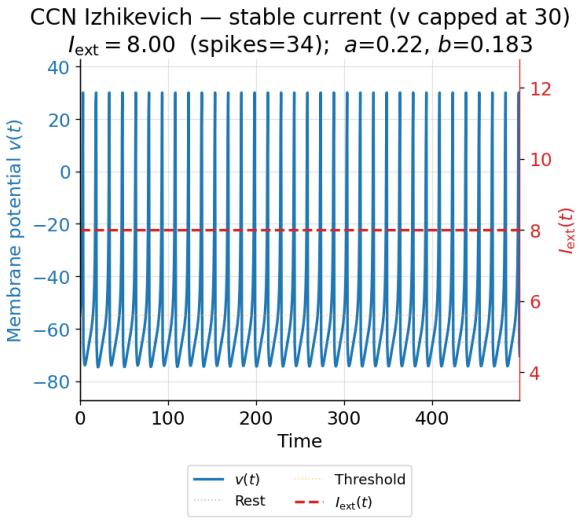


Fig. 6: Resulting spike train.

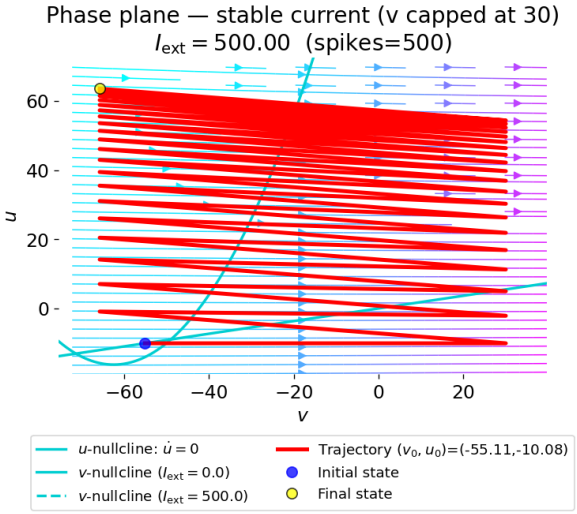


Fig. 7: When $I_{ext} = 500.0$, the system is no longer able to sustain the absolute refractory period due to hard reset values. The time period used in the graph was shortened for better readability, but the simulation ran over 500 units of time, producing 500 spikes.

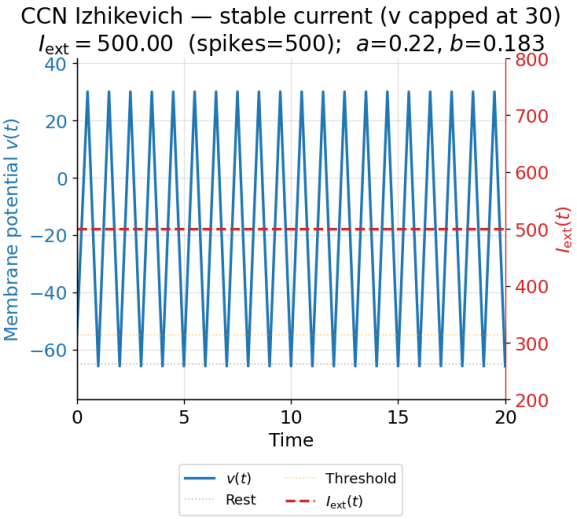


Fig. 8: Resulting spike train that no longer accommodates the absolute refractory period, as there is a spike occurring every unit of time due to the hard reset of the Izhikevich model.

3.1.2 Single Pulsed Current - Timing Dependence

Before the pulse, the trajectory sits at the $I = 0$ nullcline. During the pulse, the effective v-nullcline shifts upward by $I = 22$, which was the chosen height of the spike. As there is no random or other noise introduced to the Izhikevich neuron, the timing of the spike does not matter. The system will always follow the same trajectory and will always result in a spike of the same qualitative character.

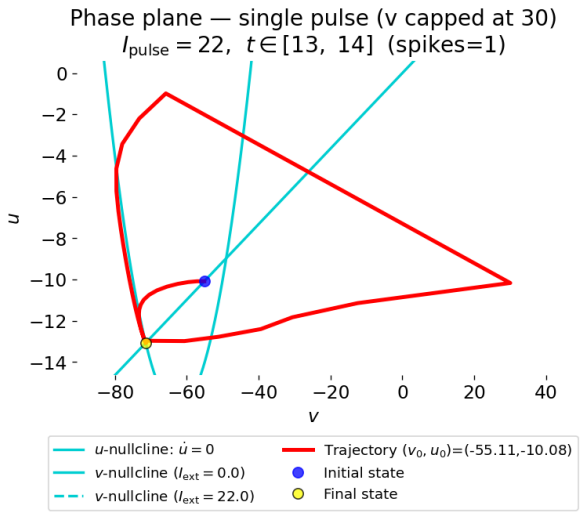


Fig. 9: Singular pulse at a time [13,14]. The pulse resulted in an action potential, as seen in a large enough dislocation of the system from the equilibrium point.

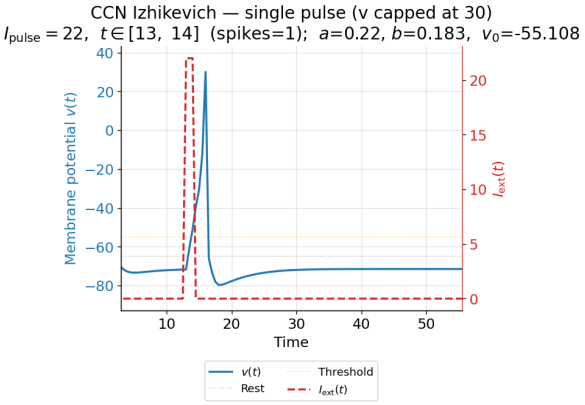


Fig. 10: The resulting action potential from a single pulse of height $I = 22$, at time [13,14].

3.1.3 Single Pulsed Current - Amplitude Dependence

With the pulse window fixed at $t \in [13, 14]$, the amplitude was varied: $I_{\text{pulse}} \in (4, 24)$, all other parameters were unchanged. For $I_{\text{pulse}} \leq 14$, no action potential occurred (*spikes* $= 0$): $v(t)$ depolarised during the pulse and relaxed afterwards, with peak deflection increasing with amplitude but remaining below the firing threshold (≈ -55 mV). For $I_{\text{pulse}} \geq 16$, exactly one spike was elicited (*spikes* $= 1$); further increases in amplitude did not change spike count or peak height (all-or-none behaviour), though depolarisation during the pulse became faster. The pulse rheobase therefore lies between 14 and 16, above the steady-state rheobase ($I_{\text{rheo}} \approx 5.0$), because a brief pulse delivers less effective charge than a sustained current of the same amplitude.

Stable current applies the same I_{ext} at every step, so sub-rheobase levels can accumulate depolarisation over time, whereas a 1-unit pulse must evoke threshold crossing before the stimulus ends and the nullcline returns, requiring a substantially larger amplitude.

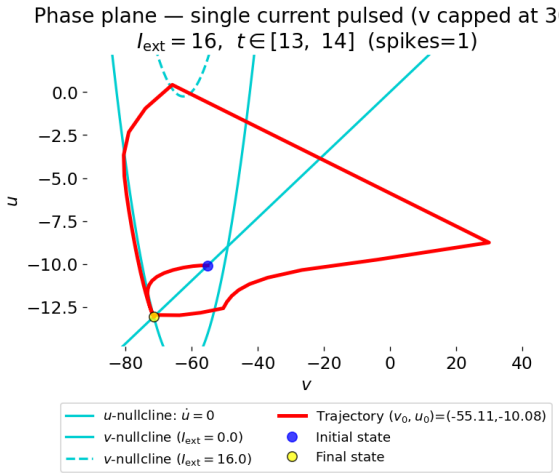


Fig. 11: As the spike was introduced only for a short unit of time, it did not result in a spike train, as it would have if it were introduced permanently.

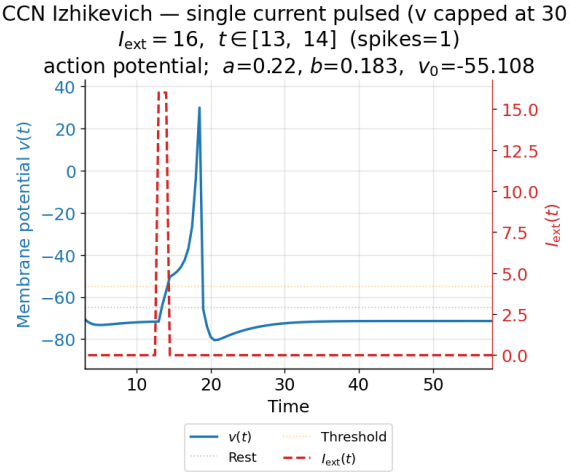


Fig. 12: Exactly one action potential occurred in spite of the height of the spike.

3.1.4 Double Pulsed Current

For double pulse stimulation, refractory effects were clearly observed. Closely spaced pulses resulted in a reduced or absent second spike, while sufficiently separated pulses produced two distinct action potentials. This demonstrates the presence of both relative and absolute refractory dynamics, where excitability is temporarily suppressed following a spike (Gazzaniga et al., 2018).

These results mirror the refractory behaviour de-

scribed for biological neurons (Gazzaniga et al., 2018): excitability is transiently reduced after a spike, and only sufficiently separated stimuli elicit a second response.

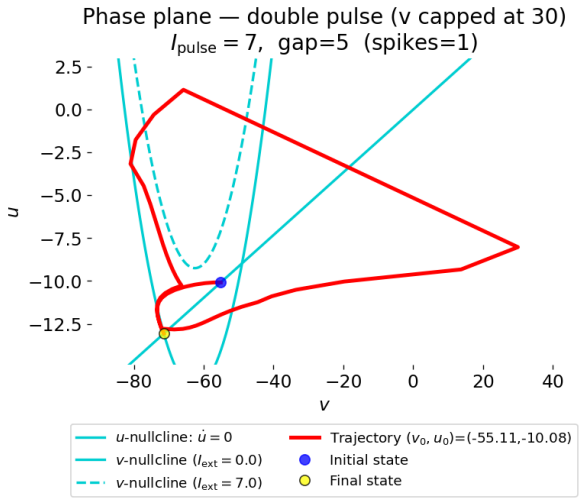


Fig. 13: Due to the absolute refractory period, no second spike occurred.

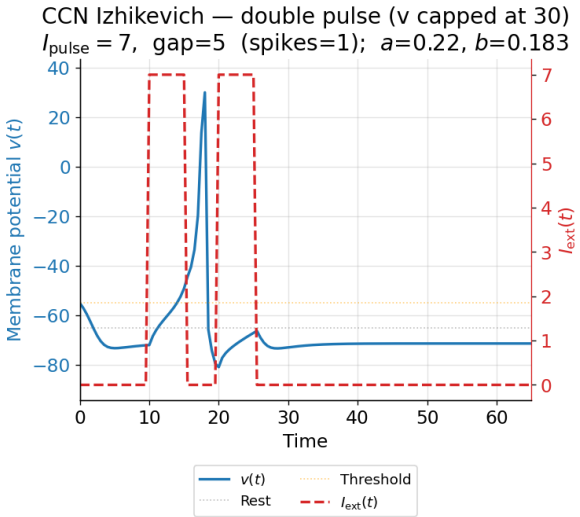


Fig. 14: Resulting voltage curve of one spike and one small oscillation.

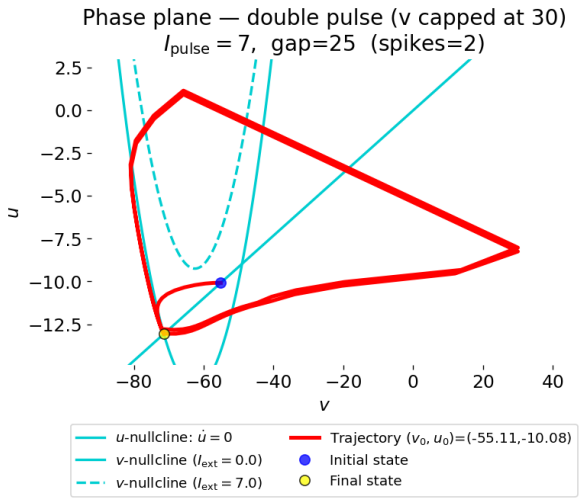


Fig. 15: As both the absolute and relative refractory periods had passed, the neuron was capable of producing two full spikes.

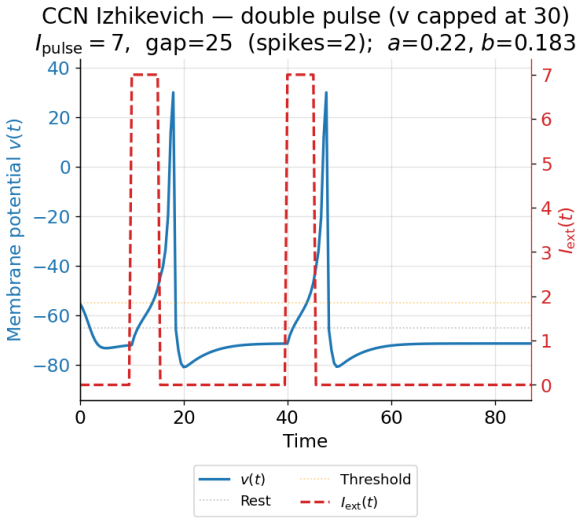


Fig. 16: Two fully pronounced spikes as a result of the two pulses with a large enough distance.

4 Conclusion

The presented analysis explored the qualitative behaviour of the Izhikevich neuron model under constant and pulsed stimulation using phase-plane methods. Across all experimental protocols, the model produces consistent excitability phenomena that align with known neuronal behaviour, including threshold-dependent firing, state-dependent responsiveness, and refractory effects.

The results confirm that these behaviours emerge directly from the interaction between the fast membrane variable v, the slower recovery variable u, and the discrete reset mechanism. In particular, the separation of

timescales between v and u is sufficient to generate rich dynamical structure without requiring detailed biophysical modelling.

Limitations

Despite its ability to reproduce a wide range of qualitative firing patterns, the model has several important limitations. First, the system operates in dimensionless units, which limits direct physiological interpretation and requires additional scaling for quantitative comparison with biological data. Second, the reset mechanism introduces a hard discontinuity into the dynamics, replacing the continuous repolarisation process observed in real neurons with an instantaneous state update. While computationally efficient, this simplification affects the realism of spike shape and recovery dynamics.

Future work

Future work will focus on a systematic exploration of the parameter space (a, b, c, d) to characterise transitions between excitability regimes in a more structured way. This would allow a more precise classification of dynamical behaviours beyond the specific parameter set used in this study. Specifically, exploration of parameter combinations that result in a behaviour consistent with the biological neuron will be searched for.

Final Evaluation

Overall, the Izhikevich model provides a computationally efficient framework for studying qualitative neuronal excitability. While it does not aim to reproduce detailed biophysical processes, it captures the essential dynamical structure underlying spike generation and adaptation. The phase-plane formulation offers a clear geometric interpretation of these behaviours. The Izhikevich model under the specific parameters introduced in this paper should be interpreted as a phenomenological tool rather than a physiologically precise description.

Acknowledgement

The authors would like to thank Prof. RNDr. Ľubica Beňušková, PhD. for her support with this research. This research was supported by SSKV (Slovak Society for Cognitive Science).

References

Cessac, B. and Samuelides, M. (2007). From neuron to neural networks dynamics. *The European Physical Journal Special Topics*, 151:135–145.

Gazzaniga, M. S., Ivry, R. B. and Mangun, G. R. (2018). *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton & Company, New York, 5th. ed.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, 14(6):1569–1572.

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press, Cambridge, MA.

Kollárová, H. (2025). Exploring the parametric sensitivity of the fitzhugh-nagumo model. *University College Roosevelt, Utrecht University*.

Musacchio, F. (2025a). Fabrizio musacchio – personal website. `https://www.fabriziomusacchio.com/`.

Musacchio, F. (2025b). Phase plane analysis. `https://github.com/FabrizioMusacchio/phase_plane_analysis`.

Dynamický Turingův test: Adaptace rozhodovacích strategií při detekci impresionistických maleb generovaných umělou inteligencí

Nikita Kostka, Ondřej Vadinský[0000−0002−0910−3140]

Katedra informačního a znalostního inženýrství, Vysoká škola ekonomická v Praze
Náměstí Winstona Churchilla 4, 130 67 Praha 3, Czech Republic
Email: nikita.kostka@gmail.com

Abstrakt

Příspěvek navrhuje dynamickou variantu grafického Turingova Testu, která umožňuje sledovat adaptaci respondentů mezi třemi koly na základě zpětné vazby. V úvodním experimentálním šetření byla sledována schopnost vysokoškolských studentů ($n = 9$) rozlišovat mezi impresionistickými malbami a obrazy generovanými modelem Midjourney (3×5 dvojic). Výsledky ukazují statisticky slabě signifikantní nárůst úspěšnosti mezi prvním a třetím kolem (z 57,7 % na 84,4 %). Pro zhodnocení psychometrických kvalit testu byla využita analýza dle principů teorie odpovědi na položku. Kvalitativní analýza odhalila tři dominantní strategie: posuzování vizuální složitosti (úspěšnost 84,5 %), řízení se uměleckým principem (úspěšnost 84,6 %) a zaměření na detaily (úspěšnost 51,6 %).

1 Úvod

Generativní umělá inteligence dnes dokáže stále věrohodněji napodobovat lidskou tvorbu, a proto je důležité zkoumat, nakolik ji lidé ještě zvládají rozpoznat. Problematikou rozlišení mezi člověkem a strojem se zabýval již britský matematik Alan Turing (1950), který navrhl experiment umožňující tuto skutečnost testovat. Tento experiment, později nazvaný Turingův test, však vznikal v době, kdy byl vývoj umělé inteligence teprve v počátcích. V kontextu dnešních technologií, zejména multimodálních modelů viz např. (Ramesh a spol., 2022; Schellaert a spol., 2023), není již původní čistě textová podoba testu postačující. Již Harnad (1989) a po něm Schweizer (2012) navrhovali rozšíření testu, aby zahrnul plné spektrum (lidského, resp. druhově specifického) inteligentního chování, byť i jim stále šlo hlavně o lepší uchopení porozumnění jazyku. Pro tento příspěvek je pak významnou inspirací především převedení Turingova testu do vizuální domény, které navrhl Michael McGuigan (2006). Turingův test i jeho dosavadní modifikace však uvažují statické měření úspěšnosti a opomíjejí proces učení, který je také podstatnou částí (nejen) lidské kognice. Předložený příspěvek proto rozšiřuje Turingův test o dynamický princip, a přenáší jej z původní textové domény do domény obrazové.

Dále příspěvek prezentuje výsledky úvodního experimentálního ověření navrženého testu. Toto ověření sledovuje dva hlavní cíle: validaci samotného metodického návrhu a následný sběr i vyhodnocení kvalitativních dat od respondentů. Právě analýza kvalitativních výsledků umožňuje identifikovat a komparovat rozhodovací strategie účastníků v jednotlivých fázích testu. Tento přístup otevírá prostor pro detailní zkoumání procesu adaptace respondentů a sledování dynamických změn v jejich úsudku v průběhu celého experimentu.

Příspěvek nejprve v sekci 2 vysvětlí teoretická východiska. Proces návrhu dynamického Turingova testu pro rozpoznávání generovaných impresionistických maleb představí sekce 3. Provedený demonstrační experiment popíše sekce 4. Jeho výsledky ukáže sekce 5. Diskuzi těchto výsledků se bude věnovat sekce 6. Příspěvek bude zakončen v sekci 7.

2 Teoretická východiska návrhu

Sekce shrnuje hlavní oblasti, na jejichž základě byl rozšířený Turingův test navržen. Technologický základ pro generování věrohodných obrazových stimulů tvoří modely generativní umělé inteligence založené na principu difúze, konkrétně model Midjourney. Teoretický základ představuje samotný Turingův test a jeho limitace, přičemž východiskem úprav pro sledování kognitivní adaptace jsou principy dynamického testování.

Difúzní modely stručně představí sekce 2.1. Turingovu testu a jeho limitům se bude věnovat sekce 2.2. Principy dynamických testů vysvětlí sekce 2.3.

2.1 Difúzní modely

Difúzní modely představují v současnosti jednu z nejprogresivnějších architektur v oblasti generativní umělé inteligence zaměřené na tvorbu obrazových dat. Jejich princip vychází z pravděpodobnostního modelování a teorie stochastických diferenciálních rovnic. Model se v rámci trénování učí dvěma základním procesům: dopředné difúzi, která postupně transformuje reálná data na šum, a následnému reverznímu procesu, při kterém dochází k rekonstrukci původních dat z náhodného šumu (Park a spol., 2023).

Proces postupné degradace obrazu (zašumování) lze v diskrétním čase $t \in \{1, \ldots, T\}$ popsat rovnicí:

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \epsilon_t \tag{1}$$

kde ϵ_t představuje nezávislé standardní Gaussovo rozdělení a skalární parametr $\beta_t \in (0, 1)$ určuje množství šumu přidaného v každém kroku. Pro další modelování se zavádí proměnná $\alpha_t = 1-\beta_t$, vyjadřující míru zachované informace. Kumulativní vliv procesu od počátečního stavu x_0 do času t je pak definován jako:

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \tag{2}$$

Cílem trénovaného modelu je predikovat přidaný šum a umožnit tak inverzní transformaci. Odšumovací krok, který směřuje od čistého šumu x_T zpět k realistickému výstupu x_0, je definován vztahem:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_q(t) z \tag{3}$$

kde $\epsilon_\theta(x_t, t)$ je neuronová síť odhadující šum v daném kroku a $\sigma_t z$ představuje stochastickou složku zajišťující rozmanitost generovaných výstupů. Pro detailnější rozbor difúzních modelů viz Higham a spol. (2023), či Ho a spol. (2020).

2.2 Turingův test

Turingův test, který je hlavní inspirací a zdrojem pro tento příspěvek, navrhl v roce 1950 britský matematik a logik Alan M. Turing ve svém článku *Computing machinery and intelligence* publikovaném v časopise *Mind*. Při experimentu jsou myšleni tři účastníci: A, B, C a hrají tzv. *imitační hru*. Jeden ze dvojice A, B je člověk, zatímco druhý je stroj. Oba komunikují s vyšetřovatelem C výhradně prostřednictvím textových zpráv (aby nebyl rozpoznatelný jejich vzhled ani chování) a snaží se ho přesvědčit, že právě oni jsou člověkem. Úkolem vyšetřovatele je na základě konverzace určit, kdo z účastníků je člověk a kdo stroj. To má vyšetřovatel určit tak, že se účastníkům A a B zadává dotazy a hodnotí jejich odpovědi dle toho, zda by takto dokázal odpovědět člověk (Turing, 1950).

Hlavní limit takto navrženého testu spočívá v jeho behaviorálním redukcionismu, kdy je inteligence posuzována výhradně na základě vnějších projevů v izolovaném prostředí, bez ohledu na vnitřní mentální procesy. Jak ilustruje Searlův experiment *Čínský pokoj*, úspěch v testu může reflektovat pouze syntaktickou dovednost manipulace se symboly, nikoliv skutečné sémantické porozumění nebo vědomí (Searle, 1980). Pro hlubší diskuzi filosofických aspektů viz např. Tvrdý (2014). Dalším kritickým faktorem je subjektivita lidského pozorovatele a binární povaha výsledku, která neumožňuje hlubší analýzu kognitivních strategií, jimiž systém k odpovědím dospěl (French, 2000). Pro diskuzi limitů Turingova testu jakožto testu (strojové) inteligence viz např. Hernández-Orallo (2000).

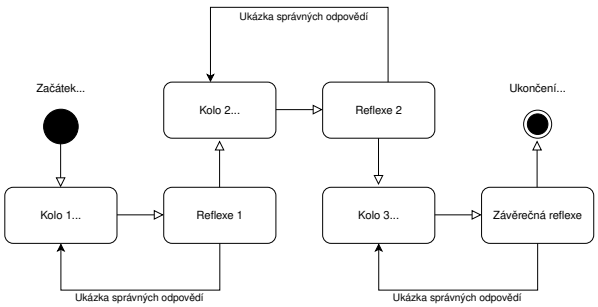
2.3 Dynamické testování

Dynamické testování představuje diagnostický přístup zaměřený na proces učení a kognitivní přizpůsobivost respondenta v čase. Na rozdíl od tradičních statických metod, které měří aktuální úroveň schopností jako neměnný stav, dynamické pojetí vychází z předpokladu, že se lidský výkon vyvíjí v interakci s prostředím a skrze poskytovanou intervenci. Klíčovým prvkem je interaktivní charakter experimentu: respondent dostává cílené instrukce, nápovědy či zpětnou vazbu, což mu umožňuje aktivně zdokonalovat své strategie a nově nabyté poznatky okamžitě aplikovat v opakovaných fázích testu (Grigorenko, 2009; Lidz, 2017). Pro širší představení dynamického testování viz Sternberg (2002).

3 Návrh dynamického Turingova testu

Tato sekce přechází od teoretických východisek k praktickému návrhu dynamického Turingova testu v doméně vizuálního umění. Hlavním cílem je vytvoření experimentálního rámce pro identifikaci a analýzu procesu kognitivní adaptace respondentů při rozpoznávání generovaných impresionistických maleb. Text popisuje celkovou strukturu testu a vysvětluje proces konstrukce jednotlivých testových položek.

Návrh předloženého dynamického Turingova test představí sekce 3.1. Konstrukci položek se bude věnovat sekce 3.2.



Obr. 1: Diagram průběhu testu

3.1 Struktura testu

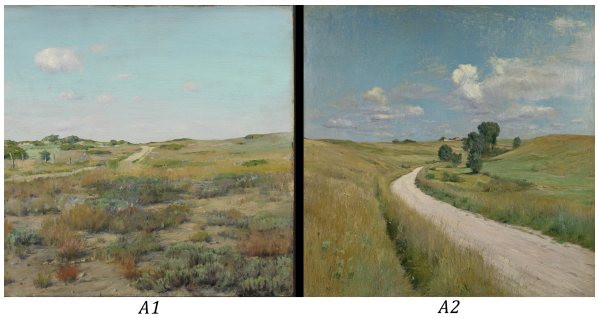
Navržená struktura přímo reflektuje principy dynamického testování popsané v sekci 2.3. Test je koncipován jako cyklický proces umožňující kognitivní adaptaci respondentů skrze průběžnou zpětnou vazbu. Celý experiment sestává ze tří kol, z nichž každé obsahuje pět položek. Jednu položku tvoří dvojice reálné impre-

sionistické malby a obrazu generovaného AI (pro bližší představení konsturkce položek viz sekci 3.2). Respondent je vyzván, aby v každé dvojici identifikoval generovaný obraz, popsal důvody své volby a zároveň uvedl hodnotu jistoty od 1 do 5 (kde 1 je nejnižší jistota a 5 nejvyšší). Zároveň je respondentovi měřen čas výběru. Všechny tyto hodnoty jsou následně zaznamenávány. Po dokončení každé pětice položek následuje fáze reflexe, v níž jsou respondentovi odhaleny správné odpovědi, čímž získává prostor pro korelaci svých poznatků a jejich aplikaci v následujících blocích.

3.2 Konstrukce položek testu

Metodika konstrukce testových položek spočívala v systematickém párování originálních děl s jejich generovanými ekvivalenty. Z digitálních archivů WikiArt [1] a Wikimedia Commons [2] bylo náhodně vybráno 15 reprezentativních impresionistických děl, každé od odlišného autora.

K tvorbě obrazů byl využit nástroj Midjourney [3]. Proces generování využíval kombinaci obrazové reference a textového popisu. Původní dílo bylo nahráno jako stylová reference, k níž byl připojen anglický prompt uvozený termínem *impressionism* a následovaný stručným popisem hlavního motivu malby.



A1 *A2*

Obr. 2: Ukázková položka testu: A1 – reálné dílo, A2 – generovaný obraz

Na Obr. 2 je zachycena ukázková dvojice: reálný obraz *William Merritt Chase – „Shinnecock Hills“* (A1) a jeho generovaný protějšek (A2) vytvořený na základě promptu *impressionism, road in fields of dry grass*.

4 Demonstrační experiment

Výše představený návrh testu byl následně podroben úvodnímu experimentálnímu ověření. Sběr dat probíhal v říjnu 2025 a testování se zúčastnilo 9 respondentů.

[1] Archiv je přístupný na `https://www.wikiart.org`.

[2] Archiv je přístupný na `https://commons.wikimedia.org`.

[3] Nástroj je přístupný skrze rozhraní na `https://www.midjourney.com`.

Respondenty blíže představí sekce 4.1. Sekce 4.2 upřesní nástroje, jenž byly při testování použité.

4.1 Respondenti

Respondenty tvořilo devět studentů z osmi českých vysokých škol v různých úrovních studia a studujících různé obory. Byli zastoupeni respondenti s odlišnými technickými znalostmi a dovednostmi, s různými zkušenostmi a postoji k využívání generativní umělé inteligence a také s rozdílnými vztahy k výtvarnému umění.

Aby byly podmínky testu metodické a srovnatelné, dodržoval výzkum doporučení Hendl (2005); Urbánek (2011) týkající se etických, administrativních a organizačních zásad. Mezi tyto principy patřilo zajištění standardizovaných podmínek testování, informovaný souhlas respondentů, srozumitelnost zadání, neutrální a objektivní pokyny a respekt k individualitě respondentů.

4.2 Použité nástroje při testování

Sběr dat probíhal formou audiozáznamu, kvantitativní tabulky a doplňujících textových poznámek administrátora. K realizaci byla využita dvě mobilní zařízení: notebook administrátora pro průběžnou evidenci údajů a měření času a tablet pro respondenta, který sloužil k interaktivní prezentaci obrazových stimulů a pořízení zvukové nahrávky. Jednotlivé testové položky byly respondentům prezentovány ve formě obrázků s možností přiblížení pro lepší analýzu detailů.

5 Výsledky

Sekce představí výsledky získané v rámci demonstračního experimentu. Kvantitativní výsledky byly analyzovány za pomocí statistických metod, kvalitativní výsledky za pomocí kódování jednotlivých odpovědí v prostředí programu QDA Miner.

Kvantitativní výsledky představí sekce 5.1. Kvantitativní výsledky představí sekce 5.2. Bližší pohled na vývoj a použití strategií mezi koly ukáže sekce 5.3. Na závěr sekce 5.4 představí provedenou analýzu teorie odpovědi na položku.

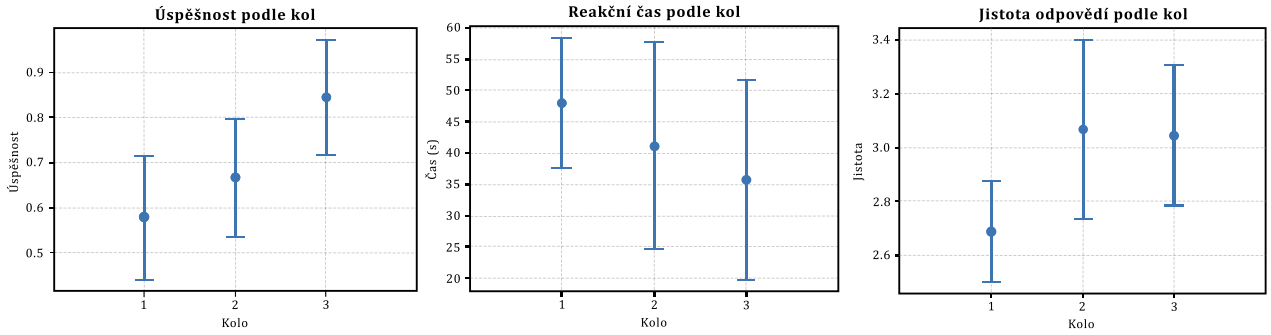
5.1 Kvantitativní výsledky

V rámci kvantitativních dat bylo nasbíráno 405 údajů ve třech sledovaných proměnných:

Správnost: booleovská hodnota, zda respondent vybral správně generovaný obrázek;

Čas: délka výběru v sekundách;

Jistota: subjektivní míra jistoty na stupnici od jedné do pěti (1 = nejmenší jistota výběru; 5 = největší).



Obr. 3: Průměrné hodnoty úspěšnosti, reakčního času a subjektivní jistoty odpovědí v jednotlivých kolech testu včetně 95% intervalů spolehlivosti.

Proměnná	Kolo 1	Kolo 2	Kolo 3
Průměrná správnost (%)	57,70	66,66	84,44
Průměrná jistota	2,68	3,06	3,04
Průměrný čas (s)	47,95	41,17	35,68

Tab. 1: Průměrné hodnoty proměnných v jednotlivých kolech.

Jak lze pozorovat ze zprůměrovaných dat v Tab. 1, průměrná správnost výběru se mezi koly postupně zvyšovala, zatímco průměrný čas výběru se snižoval. Je zde tedy přítomná jistá tendence ke zlepšování, ovšem tu se podařilo statisticky prokázat jen částečně. Jediné signifikantní zlepšení bylo v úspěšnosti mezi prvním a třetím kolem, kde dosáhl párový t-test výsledku $p = 0,0222$. Průměrná jistota mezi prvním a druhým kolem stoupla a poté se spíše ustálila.

Párový t-test: úspěšnost	p hodnota
1 a 2 kolo	p = 0.2721
2 a 3 kolo	p = 0.0688
1 a 3 kolo	p = 0.0222
Párový t-test: čas	**p hodnota**
1 a 2 kolo	p = 0.3747
2 a 3 kolo	p = 0.2454
1 a 3 kolo	p = 0.1588
Párový t-test: jistota	**p hodnota**
1 a 2 kolo	p = 0.0818
2 a 3 kolo	p = 0.8487
1 a 3 kolo	p = 0.0864

Tab. 2: P hodnoty porovnání jednotlivých kol.

Pro lepší představu jsou výsledky vyznačeny v grafech s 95% intervalem spolehlivosti na Obr. 3. Velký rozsah intervalů spolehlivosti je dán nízkým počtem subjektů ($n = 9$) a v některých případech zřejmě i vyšší variabilitou v jejich odpovědích. Pozorovatelné překryvy intervalů spolehlivosti naznačují, že ve většině případů nejsou průměrné hodnoty sledovaných proměnných signifikantně odlišné, což je v souladu s výsledky provedeného párového t-testem.

5.2 Kvalitativní výsledky

Jak již bylo zmíněno v úvodu této sekce, kvalitativní výsledky byly analyzovány za pomocí metody kvalitativního kódování. Analýza probíhala induktivním způsobem, kdy byly kódy generovány přímo na základě dat, bez předem stanoveného teoretického rámce. V první fázi byly seskupeny poznámky získané při pozorování jednotlivých respondentů a segmentům textu s použitou strategií byly přiděleny popisné kódy zachycující hlavní motiv. Dále byly tyto kódy roztříděny aby mezi nimi bylo možné najít kauzální vztahy. Na závěr byly dílčí kódy seskupeny do širších tématických celků, u nichž byly dále sledovány změny výskytů v jednotlivých kolech, které umožnily vyvození hlavních závěrů. Výsledné rozdělení ukazuje Tab. 3.

Kód	Počet výskytů	Úspěšné určení	Úspěšnost (%)
Plátno	5	3	84,5
Detaily	31	16	51,6
Složitost	58	49	84,5
Odraz vody	11	6	54,5
Pocitově	21	11	52,4
Umělecký princip	26	22	84,6
Zkušenosti s AI	6	4	66,7
Bez uvedení důvodu	17	12	70,6

Tab. 3: Kvalitaitvní výsledky – kódy strategií.

Již na první pohled je zřejmá nějčastěji používaná strategie, kterou je pozorování nepřirozené složitosti generovaných obrazů. Tato strategie také patří k jedněm z nejúspěšnějších. Dalšími dvěmi nejčastěji použitými strategiemi je pozorování detailů a analýza obrazů na základě uměleckého principu.

5.3 Vývoj strategií mezi koly

Mnohem zajímavější je ovšem pohled na celkový vývoj použití jednotlivých strategií a jejich aplikaci následnující reflexi předešlého kola. Tam je nejvíce patrné, jak v závislosti na zpětné vazbě dokázali respondenti upravovat použití jednotlivých strategií. Níže je rozebrán vývoj tří nejčastějších strategií představených v minulé sekci.

Nejlepší ukázkou správné adaptace je vývoj strategie pozorování umělé složitosti generovaných maleb. Ta byla v prvním kole použita ve 12 případech, ovšem pouze u 6 z nich vedla k úspěšnému rozlišení. Dosáhla tedy úspěšnosti 50 %, což je stejné jako náhodné hádání. Poté proběhla první reflexe a respondenti měli možnost zhodnotit své dosavadní výsledky. Ve druhém kole byla tato strategie využita již v 17 případech, ze kterých v 16 případech pomohla ke korektnímu určení a dosáhla tedy úspěšnosti 94 %. Je tedy vidět relativně slabý nárůst v množství výskytů, ale znatelný nárůst ve správnosti určení. Ve třetím kole byla poté tato strategie využita ve 29 případech, z čehož ve 27 vedla ke správnému určení a dosáhla tak úspěšnosti 93 %. Lze tak pozorovat, že zatímco úspěšnost zůstala stále extrémně vysoká, značně se také zvýšilo nasazení této strategie. Vývoj využití této strategie by tak mohl naznačovat, že ji účastníci po první reflexi začali vnímat jako účinnější a začali na ni více spoléhat. Značné zlepšení mezi prvním a druhým kolem také může značit, že se respondenti naučili používat ji efektivněji v určitých případech.

Opačný případ lze pozorovat u strategie pozorování detailů. Tato strategie byla v prvním kole nasazena v 17 případech a k úspěšnému odhalení generované malby vedla v 11 z nich. Dosáhla tedy 64% úspěšnosti. Ve druhém kole byla využita v 11 případech, ovšem pouze ve 4 z nich úspěšně a dosáhla tak úspěšnosti pouze 36 %. Lze tedy pozorovat, že se samotné nasazování strategie snížilo, ale ještě více se snížila její úspěšnost. Ve třetím kole byla využita již pouze ve třech případech a pouze v jednom z nich vedla ke správnému určení. Tento vývoj naznačuje, že v průběhu testu začali respondenti upřednostňovat jiné a účinější strategie a od této pomalu upouštěli.

Zajímavé také je, že v prvním kole se jednalo zcela o nejpoužívanější strategii, což poukazuje na obecný předsudek respondentů vůči umělé inteligenci, že chyby v generování jsou nejlépe patrné v chybách detailů. Nízká úspěšnost strategie ovšem může znamenat, že generativní modely již neprodukují zjevné chyby v detailech, které by bylo možné snadno odhalit, či že se nejedná o vhodný znak generování specificky u impresionistických obrazů, kde umělci často nevěnovali detailům přílišnou pozornost a snažili se hlavně o působení obrazu jako celku.

Třetí nejčastější strategie, tedy řízení se uměleckým principem má vývoj relativně stálý. V prvním kole byla tato strategie použita v 8 případech, z čehož v 6 efektivně. Měla tedy 75% úspěšnost. Ve druhém kole byla využita opět v 8 případech, ale v 7 vedla ke správnému určení. Ve třetím kole byla využita v 10 případech a v 9 z nich efektivně. Lze tedy pozorovat relativně stálý počet aplikace strategie a také stálou, relativně vysokou úspěšnost. Tento výsledek by mohl značit, že se jedná o spolehlivou strategii nejlépe aplikovatelnou u určitých položek, kde jsou patrné stylové prvky.

Tato analýza jednotlivých strategií ukazuje, že zaměření na změny mezi koly testu přináší nový a hlubší pohled do adaptací jednotlivých respondentů a mělo by smysl je sledovat více. Ovšem je důležité upozornit, že v rámci testování nebyl pro respondenty limitován počet strategií při výběru jedné položky. Tedy u jednoho rozhodnutí mohlo být použito více strategií zároveň.

5.4 Teorie odpovědi na položku

Pro zhodnocení položek testu byla využita analýza za pomoci teorie odpovědi na položku, jak popisují Allen a Yen (2001). Ta odhalila značné limity v některých položkách i struktuře testu. Výsledky ukazují, že test byl pro respondenty spíše jednodušší (průměrná p-hodnota 0,69) a celková průměrná diskriminační korelace (0,092) signalizuje nižší strukturní stabilitu testu. Ovšem interpretace obtížnosti je částečně ovlivněna fixním řazením položek. Vysoká úspěšnost v pozdějších fázích může být důsledkem procesu učení, což je v souladu s dynamickou povahou testu.

Podařilo se však určit položky jak se silnou, tak i se slabou výpovědní hodnotou. U některých položek byla korelace záporná, což značí nekonzistenci s celkovým výkonem. Slabší respondenti na ně tedy odpovídali lépe něž ti silnější.

Analýza IRT tak poskytuje podklad pro budoucí optimalizaci výběru a pořadí stimulů.

6 Diskuse výsledků

Tato sekce interpretuje získaná data, hodnotí metodické limity testu a navrhuje směry pro navazující výzkum. Výsledky jsou nejdříve zasazeny do širšího kontextu v sekci 6.1. Kritická reflexe limitů následuje v sekci 6.2. Možnosti dalšího rozvoje rozebírá sekce 6.3.

6.1 Interpretace hlavních zjištění

Z kvantitativních výsledků se statisticky podložené zlepšení projevilo pouze u úspěšnosti v porovnání prvního a třetího kola. U zbývajících hodnot, jako je průměrný čas výběru a subjektivní jistota, nelze kvůli nízkému počtu participantů činit definitivní závěry. Přesto lze mezi jednotlivými koly pozorovat tendenci ke snižování časové latence a postupné stabilizaci jistoty respondentů. Kvalitativní analýza naopak jasně ukázala, že se v průběhu testu stala dominantní strategie zaměření na umělou složitost obrazu, přičemž respondenti v její prospěch opouštěli méně účinné postupy. Spojením těchto zjištění lze vyvodit, že respondenti v průběhu testu aktivně vyhodnocují své strategie, což jim pomáhá k úspěšnějšímu rozlišování mezi reálnou a generovanou tvorbou. Návrh testu tak i přes své limity nabízí funkční metodický rámec pro sledování kognitivní adaptace.

6.2 Reflexe a limity

Navržený test má v nynější podobě řadu limitů. Nejzásadnějším faktorem je malý počet respondentů ($n = 9$), což je způsobeno časovou náročností hloubkových rozhovorů. Při větším vzorku by bylo možné dosáhnout robustnějších výsledků vhodnějších pro statistické ověření. Další limity vyplývají z analýzy IRT, která odhalila nevyrovnanou obtížnost a diskriminační schopnost některých testových dvojic, což snižuje ucelenost testu.

Zásadním omezením je také výhradní zaměření na impresionistické malby. Toto rozhodnutí sice zjednodušilo analýzu, ale limituje generalizaci výsledků na jiné umělecké styly či doménu fotografie. Identifikovaná strategie pozorování složitosti může být specifická právě pro impresionismus, kde modely generativní AI zatím nedokáží plně napodobit charakteristický rukopis. V neposlední řadě představuje limit i neustálý technický vývoj AI. Je pravděpodobné, že s nástupem dokonalejších modelů budou současné identifikační strategie vyžadovat revizi.

6.3 Možnosti dalšího výzkumu

Budoucí výzkum by se měl zaměřit na eliminaci zmíněných limitů. Přirozenou cestou je aplikace testu na další umělecké směry, což umožní analyzovat proměnlivost strategií v závislosti na vizuálním charakteru děl. Zajímavý vhled by mohlo přinést také zkoumání vlivu informovanosti respondenta, tedy zda a jak se změní úspěšnost výběru, pokud je účastníkům předem vysvětlen princip fungování generativních modelů.

Zásadní prostor pro rozvoj představuje hlubší analýza individuálních strategií. Dosavadní data naznačují, že detailní sledování toho, jak se konkrétní strategie u jednotlivých respondentů v čase utvrzují či mění, nabízí nejvyšší výpovědní hodnotu o procesu lidského učení. V metodologické rovině se dále nabízí randomizace pořadí položek pro eliminaci únavového efektu nebo vytvoření různých položek, které budou nejprve samostatně zhodnoceny pro co nejvyšší rozlišovací schopnost.

V neposlední řadě by bylo přínosné srovnání výkonu laiků s experty z oblasti dějin umění a vývojářů AI, kteří by mohli k rozlišování přistupovat na základě odlišných zdůvodnění výběru.

7 Závěr

Tento příspěvek představil modifikaci Turingova testu převedením do obrazové domény a jeho rozšířením o dynamický princip. Tato úprava umožnila sledovat změny v rozhodovacím procesu respondentů a analyzovat, jakým způsobem přizpůsobují své strategie na základě získané zpětné vazby.

Kvantitativní data potvrdila statisticky významné zlepšení úspěšnosti mezi počátkem a koncem testu. Kvalitativní analýza identifikovala klíčové strategie, z nichž nejefektivnější se ukázalo posuzování kompoziční složitosti.

I přes popsané limity v podobě malého výzkumného souboru a úzkého stylového zaměření práce potvrzuje, že dynamický přístup k testování poskytuje hlubší vhled do interakce člověka s obsahem generovaným umělou inteligencí. Předložený koncept testu může sloužit jako základ pro další zkoumání hranic mezi lidskou a strojovou tvorbou v neustále se vyvíjejícím digitálním prostředí.

Reference

Allen, M. J. a Yen, W. M. (2001). *Introduction to Measurement Theory*. Waveland Press.

French, R. M. (2000). The Turing Test: the first 50 years. *Trends in Cognitive Sciences*, 4(3):115–122.

Grigorenko, E. L. (2009). Dynamic Assessment and Response to Intervention. *Journal of learning disabilities*, 42(2):111–132.

Harnad, S. (1989). Minds, machines and searle. *Journal of Experimental & Theoretical Artificial Intelligence*, 1(1):5–25.

Hendl, J. (2005). *Kvalitativní výzkum: základní metody a aplikace*. Portál.

Hernández-Orallo, J. (2000). Beyond the turing test. *Journal of Logic, Language and Information*, 9(4):447–466.

Higham, C. F., Higham, D. J. a Grindrod, P. (2023). Diffusion Models for Generative Artificial Intelligence: An Introduction for Applied Mathematicians.

Ho, J., Jain, A. a Abbeel, P. (2020). Denoising Diffusion Probabilistic Models.

Lidz, C. (2017). Dynamic Assessment: Thinking Assessment for the Thinking Teacher1 Carol S. Lidz. *The Thinking Teacher*, Volume 1.

McGuigan, M. (2006). Graphics Turing Test.

Park, Y.-H., Kwon, M., Choi, J., Jo, J. a Uh, Y. (2023). Understanding the Latent Space of Diffusion Models through the Lens of Riemannian Geometry.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. a Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents.

Schellaert, W., Martínez-Plumed, F., Vold, K., Burden, J., Carases, P. A. M., Loe, B. S., Reichart, R., HÉigeartaigh, S. Ó., Korhonen, A. a Hernández-Orallo, J. (2023). Your prompt is my command: On assessing the human-centred generality of multimodal models. *Journal of Artificial Intelligence Research*, 2023(77):377–394.

Schweizer, P. (2012). Externalist foundations of a truly total turing test, the. *Minds and Machines*, 22(3):191–212.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, (3):417–457.

Sternberg, R. J. (2002). *Dynamic testing : the nature and measurement of learning potential*. Cambridge, UK ; New York : Cambridge University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

Tvrdý, F. (2014). *Turingův Test: Filozofické Aspekty Umělé Inteligence*. Togga, Praha, 1. vyd.

Urbánek, T. (2011). *Psychometrika: měření v psychologii*. Portál.

Exploring the Tolman-Eichenbaum Machine for Modeling Spatial Understanding in Robots

Michal Kostrhun, Igor Farkaš

Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava
Mlynská dolina 6284, 842 48 Bratislava
kostrhun1@uniba.sk, igor.farkas@fmph.uniba.sk

Abstract

Spatial understanding is a crucial cognitive component in all agents acting in 3D physical space, including people and robots. We investigate the Tolman-Eichenbaum Machine –a unifying mechanistic framework for the hippocampal role in spatial and non-spatial tasks– for use in biologically inspired Simultaneous Localization and Mapping (SLAM) in robots. Establishing the preliminary context of neuroscience, we provide an approachable overview to the model and its capabilities. We compare it to other approaches used in previous bio-inspired SLAM systems and address their limitations. We summarize advantages and limitations of TEM and briefly outline our future work regarding its use in cognitive robotics.

1 Introduction

Navigation of robotic agents in real-world environments is a complex task that has been actively researched for decades. In particular, we focus on simultaneous localization and mapping (SLAM). The main challenge with SLAM has always been the spatio-temporal complexity of the environment. Navigation in static environments has been computationally solved in the 1990's (Oriolo et al., 1998), however as the environments get more dynamic and complex, traditional methods fall short of achieving efficient solutions. Current visual SLAM methods fail to achieve real-time performance (Sun et al., 2017) and leveraging semantic information to develop semantic SLAM remains an active area of research (Nahavandi et al., 2025).

In our work, we take a different approach. It is known that hippocampus in mammal brains is essential for survey navigation and evidence points towards its involvement in other types of navigation where location memory is essential (Eichenbaum, 2017). The Tolman-Eichenbaum Machine (TEM) is a machine learning-based computational model of hippocampus and entorhinal cortex, formally linking them together (Whittington et al., 2020). The authors demonstrated that TEM cells resemble different cell types also found in organic hippocampus (grid cells, place cells) and learns to path integrate in unknown environments.

In this article we introduce TEM in the context of neuroscience and cognitive science, going over important preliminaries to understanding the model. We further introduce extensions to the original TEM, aiming to make it suitable for use in SLAM tasks, which will also be the subject of our future work.

2 Preliminaries

The hippocampus is a seahorse-shaped structure located in temporal lobe of the brain. It receives input from the entorhinal cortex (Witter, 2011) and plays an essential role in creating cognitive maps of the environment in animals and humans (biological agents). The neuronal representations of cognitive maps are facilitated by various neuronal cell types, including, but not limited to grid cells and place cells.

Grid cells are the most abundant cell type in medial part of EC. They represent a role of a tiling, layered positioning system in the real world. A single grid cell represents a regular, hexagonal grid of fields. When a biological agent enters one of the fields in the real world, the corresponding grid cell starts firing (Rowland et al., 2016).

Place cells are located primarily in the hippocampus. Their firing signals encode a particular location in real-world environment. Due to remapping, a single place cell can encode different locations depending on the environment (Moser et al., 2015).

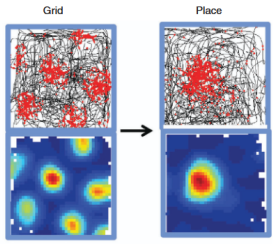


Fig. 1: Grid cell and place cell (Moser et al., 2015). **(Top)** Location of neural spikes (red) superimposed onto trajectory (black) of a biological agent. **(Bottom)** Color-coded rate map with red showing high activity and blue showing low activity.

3 Previous work in Bio-inspired SLAM

Remarkably, the neural representations and processes in the hippocampal-EC system align with the task of SLAM. A biological agent placed in a new environment organizes the perceived stimuli into a cognitive map which is then referenced when localization is performed.

3.1 Overview

We acknowledge two previous works in the field of biologically inspired SLAM models. These works are RatSLAM by Milford et al. (2004) and ViTa-SLAM (Visuo-Tactile SLAM) by Struckmeier et al. (2019). They were chosen as illustrative examples for both older and more recent methods of modeling bio-inspired SLAM.

RatSLAM implements the hippocampal processes using a competitive attractor network of "pose cells" that combine the functions of biological place and head direction cells. By mimicking rodent navigation through the dynamics of "activity packets" and using Hebbian learning for visual association, the system creates a flexible, non-Cartesian representation capable of maintaining multiple spatial hypotheses to resolve environmental ambiguity.

ViTa-SLAM implements the hippocampal-EC model by using a 3D pose cell network designed as a Continuous Attractor Network (CAN) to emulate the spatial representation of grid cells. To resolve environmental ambiguity, the system mimics the biological reliance on whiskers (vibrissae) to supplement coarse vision, fusing unique visual templates with non-unique tactile feedback. This integration utilizes active whisking strategies to enhance sensory data quality and produces a semi-metric "ViTa map", replicating the biological process of organizing disparate stimuli into a cohesive cognitive map for robust localization.

3.2 Limitations

While both of the above works demonstrate novel, bio-inspired SLAM mechanisms for real robots, several technical and practical limitations persist across these models.

Evaluation often focuses on topological consistency rather than Cartesian accuracy, which may not meet standard metric requirements. A primary bottleneck is the high computational cost of maintaining large competitive attractor networks as environments grow. Practically, the reliance on non-standard sensors, such as the use of biomimetic whiskers, limits the immediate portability of these systems to general robotic platforms.

Additionally, correcting accumulated path integration drift upon returning to a previously visited area remains difficult. This phenomenon is predominantly observed in large, circular arenas where topological relationships are harder to resolve.

The biological implementation also remains a highly simplified abstraction. RatSLAM integrates position and orientation into a single network, whereas biological agents utilize distinct populations of place cells and head-direction cells. Finally, these systems lack advanced biological mechanisms like predictive coding and active spatial attention, which are fundamental to how animals filter sensory data and adapt to new environments.

4 Tolman-Eichenbaum Machine

The Tolman-Eichenbaum Machine (TEM), introduced by Whittington et al. (2020) builds upon research advancements in the field of neuroscience. Notably, the idea that relational memory and spatial reasoning may share a common mechanism (Eichenbaum and Cohen, 2014) and that this mechanism can be modeled by finding an efficient representation of a connected graph (Stachenfeld et al., 2017).

Implementing these ideas, authors demonstrate that TEM generalizes structural knowledge in space and non-space (SNS), while also correlating with the neural phenomenons measured empirically in hippocampus of rat and mice. Remarkably, hippocampal remapping of place cells, previously thought to be random, was predicted by TEM to preserve hidden structural representations. These results suggest TEM to be a general framework for hippocampal-entorhinal representation, inference, and generalization across spatial and non-spatial tasks.

4.1 Overview

In the original paper by Whittington et al. (2020), the authors consider an unsupervised learning problem where an agent is moved across the nodes of a sensory graph, receiving each edge transition and sensory input from the current node as an input pair. A sensory graph's edges represent structured relations in SNS, such as the bonds of a family tree (Fig. 2A) or discrete, ordinal movements in euclidean space (Fig. 2B). Each node in a sensory graph represents a unique sensory information.

The TEM is designed to resemble the hippocampal-EC formation. It receives sensory input encoded by lateral entorhinal cells, which is combined with abstract location representation provided by medial entorhinal cells using a conjunctive process in TEM's hippocampus. Hebbian learning on auto-associative networks is used to model TEM's hippocampal place cells. The mechanism behind abstract location in TEM is represented by a recurrent neural network (RNN)-like model.

On a high level, TEM operates by first learning the structure of a family of sensory graphs representing a particular SNS during training. This trains TEM's structural representation, or grid cells. Then, an agent is placed on a sensory graph and TEM, running separately, receives the agent's edge transitions and sensory information from each visited node. On each step, TEM stores its representation of input sensory information ("what") in conjunction with its location representation ("where") inside its auto-associative network. This allows the future retrieval of "what" upon providing "where" (and vice versa) via attractor mechanics. TEM predicts the next "what" by inferring the next "where" from the desired transition, then retrieving the missing "what" of the conjunction with "where" using attractor mechanics. This is, of course, assuming that the node has already been visited.

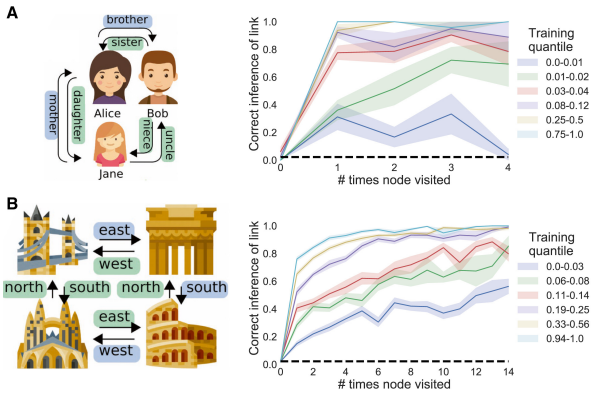


Fig. 2: Learning to learn: when TEM has only seen a few environments (blue/green) it takes many visits to each node to remember it (Whittington et al., 2020). This is because it (1) does not yet understand the structure of the graph and (2) has not learned how to use memories. After visiting more environments and learning the common structure (cyan/yellow), TEM correctly predicts a node on the second visit regardless of the edge taken—TEM now understands both the rules of the graph (path integration) and how to store and retrieve memories. **(A)** social hierarchies, and **(B)** 2D graphs.

4.2 Advantages

Path integration accumulates errors, which the TEM architecture addresses by allowing the refinement of abstract location representations through the retrieval of conjunctive memories. Sensory data acts as a landmark; the TEM indexes its hippocampal memory with current observations to retrieve candidate locations, which better estimate the real location using attractor dynamics. This mechanism provides a more biologically grounded solution for resolving environmental ambiguity than the competitive attractor dynamics used in RatSLAM and its descendants.

Furthermore, as Whittington et al. (2020) demonstrated in their original paper, the TEM architecture empirically resembles the biological hippocampal-entorhinal formation. While RatSLAM simplifies rodent navigation by integrating position and orientation into a single "pose cell" network, TEM maintains distinct populations for structural abstraction (medial entorhinal cortex) and sensory binding (hippocampus). This separation enables the emergence of diverse cell types, including grid, border, and object vector cells that reflect the transition statistics of the environment.

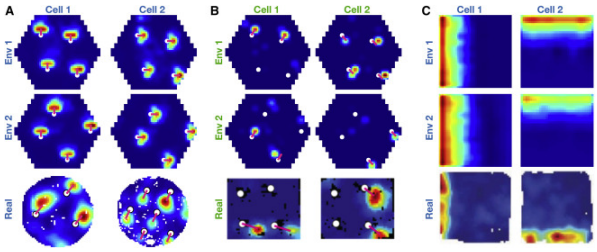


Fig. 3: TEM's various cell types compared to real data measured from rodents (Whittington et al., 2020). **(A)** Object vector cells, which generalize to any object, firing when agent is located at a constant vector offset from an object. **(B)** Landmark cells, similar to object vector cells, though only for specific objects. **(C)** Boundary cells, firing when agent is near a specific boundary.

A primary advantage for robotics is TEM's ability to generalize to novel environments. By factorizing knowledge into structural regularities ("where") and sensory content ("what"), which are separate during training, the system can transfer learned spatial rules to new sensory graphs. Once the underlying structure is learned, a robotic agent can perform first-presentation inferences, predicting transition consequences for nodes it visited through edges it has never physically traversed.

4.3 Limitations

Despite these advantages, several challenges must be addressed to make TEM a viable SLAM solution for general robotics. First, the original TEM architecture is designed to operate on discrete graphs, where movement is defined by edge transitions between adjacent nodes. Real-world robotic navigation occurs in continuous 3D Euclidean space and time, requiring the model to be extended beyond simple ordinal transitions to handle non-discrete coordinates and the full spatio-temporal complexity of robotic tasks.

Second, while TEM works with arbitrary sensory information (often represented as 1-hot vectors in simulations), robots require the processing of structured high-dimensional data, such as vision. While SLAM is most often solved using visual input, the original model's reliance on random sensory values at each node of the sensory graph does not account for the multi-

modal and complex nature of real-world stimuli. To be effective, TEM must be integrated with robust visual encoders, deriving usable and distinct features from raw pixel data into the representations required for hippocampal binding.

Finally, the computational demands of the architecture may pose a barrier to real-time performance, a limitation already observed in existing visual SLAM methods. The use of iterative attractor networks for memory retrieval and the reliance on training processes like backpropagation through time can be computationally expensive. As environments grow in complexity, maintaining the large auto-associative matrices necessary for high-capacity memory may lead to bottlenecks on mobile robotic platforms.

5 Future work

In the original article, Whittington et al. (2020) introduced the TEM and proved its capabilities in robust mapping and localization across space and non-space. Showing promising results across the various test cases designed to unambiguously test its relational memory, TEM's hippocampal and EC cells have demonstrated empirically similar behavior to real ones found in rodents. Remarkably, TEM's architecture hints at explanations of previously unexplained phenomena, such as remapping of hippocampal place cells.

While the original article focused on proving the above, little, if any, follow-up work has been done to utilize this model in robotics. In our future work, we will do just that. Addressing the above limitations, we will implement the model's architecture in a humanoid robot NICO. We will devise test cases and environments to test the model's capabilities in simulated 3D environments translatable to real-world scenarios. Upon successful implementation, we anticipate the first challenge to be the robot's dynamic peripersonal space where objects can be physically manipulated (by the robot or another agent). Lastly, we will measure our model's performance using standardized metrics and will compare our results with other, bio-inspired and state-of-the-art SLAM systems.

Acknowledgement

This work was supported by Horizon Europe MSCA project no. 101072488. M.K. acknowledges the grant from the Slovak Society for Cognitive Science.

References

Eichenbaum, H. (2017). The role of the hippocampus in navigation is memory. *Journal of Neurophysiology*, 117(4):1785–1796.

Eichenbaum, H. and Cohen, N. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron*, 83(4):764–770.

Milford, M. J., Wyeth, G. F. and Prasser, D. (2004). RatSLAM: a hippocampal model for simultaneous localization and mapping. In *IEEE International Conference on Robotics and Automation*, pp. 403–408.

Moser, M.-B., Rowland, D. C. and Moser, E. I. (2015). Place cells, grid cells, and memory. *Cold Spring Harbor perspectives in biology*, 7(2):a021808.

Nahavandi, S., Alizadehsani, R., Nahavandi, D., Mohamed, S., Mohajer, N., Rokonuzzaman, M. and Hossain, I. (2025). A comprehensive review on autonomous navigation. *ACM Computing Surveys*, 57(9):1–67.

O'Keefe, J. and Nadel, L. (1979). The cognitive map as a hippocampus. *Behavioral and Brain Sciences*, 2(4):520–533.

Oriolo, G., Ulivi, G. and Vendittelli, M. (1998). Real-time map building and navigation for autonomous robots in unknown environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3):316–333.

Rowland, D. C., Roudi, Y., Moser, M.-B. and Moser, E. I. (2016). Ten years of grid cells. *Annual review of neuroscience*, 39(1):19–40.

Stachenfeld, K. L., Botvinick, M. M. and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643–1653.

Struckmeier, O., Tiwari, K., Salman, M., Pearson, M. J. and Kyrki, V. (2019). Vita-SLAM: A bio-inspired visuo-tactile slam for navigation while interacting with aliased environments. arXiv:1906.06422.

Sun, Y., Liu, M. and Meng, M. Q.-H. (2017). Improving RGB-D SLAM in dynamic environments: A motion removal approach. *Robotics and Autonomous Systems*, 89:110–122.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N. and Behrens, T. E. (2020). The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.

Witter, M. (2011). Entorhinal cortex. *Scholarpedia*, 6(10):4380.

Analýza citlivosti genetických algoritmů na nastavení parametrů: srovnání fixního a adaptivního přístupu

Anna Krajčírová

Ústav informatiky, Slezská univerzita v Opavě
Bezručovo nám. 13, 749 01 Opava
Email: anna.krajcirova@fpf.slu.cz

Abstrakt

Genetické algoritmy (GA) představují optimalizační metodu inspirovanou evolucí. Jsou využívány při řešení složitých nebo rozsáhlých úloh, u nichž nelze efektivně aplikovat klasické postupy. Výkonnost GA je významně ovlivněna nejen formulací fitness funkce, ale také nastavením jeho parametrů. Nevhodná volba těchto parametrů může ovlivnit rychlost konvergence i kvalitu nalezeného řešení.

Cílem tohoto článku je experimentálně porovnat fixní a adaptivní řízení parametrů GA a analyzovat jejich vliv na průběh evoluce a výslednou kvalitu řešení. Srovnávací analýza je provedena na třech typech úloh: (1) maximalizaci jednorozměrné funkce, (2) maximalizaci dvourozměrné funkce a (3) řešení problému obchodního cestujícího. Na základě experimentálních výsledků jsou porovnány rozdíly mezi oběma přístupy a jejich praktické implikace.

1 Úvod

Genetické algoritmy (GA) spadají do třídy evolučních algoritmů (EA), inspirovaných principy přirozeného výběru známými z biologie. S myšlenkou evolučních výpočtů přišel v roce 1960 Ingo Rechenberg ve své práci *Evolution Strategies*. Za průkopníka genetických algoritmů je považován John Holland, který se svými studenty v raných 70. letech položil základy této oblasti (Trinh, 2025).

Genetické algoritmy jsou široce využívány při řešení optimalizačních problémů, zejména v případech, kdy je nalezení přesného řešení výpočetně náročné nebo prakticky neproveditelné. Typickými oblastmi jejich použití jsou kombinatorické úlohy, nelineární optimalizace nebo problémy s mnoha lokálními extrémy, ve kterých mohou tradiční gradientní metody selhávat.

Jednou z klíčových vlastností genetických algoritmů je schopnost prohledávat rozsáhlý stavový prostor bez nutnosti explicitní znalosti jeho struktury. Na druhou stranu je jejich výkonnost silně ovlivněna volbou parametrů, jako jsou velikost populace, pravděpodobnost křížení a pravděpodobnost mutace. Nevhodně zvolené parametry mohou vést k předčasné konvergenci, stagnaci algoritmu nebo naopak k nedostatečnému využití nalezených řešení (Le a spol., 2023).

Tradiční přístup k návrhu genetických algoritmů využívá pevně nastavené hodnoty těchto parametrů, které jsou určeny před zahájením výpočtu. Alternativou je adaptivní řízení parametrů, kdy se jejich hodnoty mění v průběhu evoluce v závislosti na aktuálním stavu hledání. Tento přístup umožňuje lépe vyvažovat fáze explorace, tedy prohledávání dosud neprozkoumaných oblastí stavového prostoru, a exploitace, tedy intenzivního prohledávání okolí již nalezených kvalitních řešení. Adaptivní přístup tak má potenciál zvýšit robustnost algoritmu.

Cílem tohoto článku je analyzovat vliv nastavení parametrů genetického algoritmu na jeho výkonnost a porovnat přístup s fixními parametry a adaptivní variantou algoritmu. Experimenty jsou provedeny na třech typech úloh s různou mírou složitosti: jednorozměrné a dvourozměrné optimalizační úloze a problému obchodního cestujícího (TSP). Výsledky jsou hodnoceny z hlediska kvality nalezeného řešení, rychlosti konvergence a stability mezi jednotlivými běhy.

2 Genetické algoritmy

Biologickou motivací genetických algorimtů je napodobování principů biologické evoluce a přirozeného výběru při řešení optimalizačních a prohledávajících úloh. Dochází k simulaci vývoje populace jedinců, která se postupně přízpůsobuje svému prostředí stejně, jako tomu je v přírodě. Jinými slovy, populace se zlepšuje v řešení daného problému.

V biologii jsou genetické informace uchovávány v chromozomech, které jsou předávány potomkům. Z informatického hlediska lze chromozom chápat jako kód reprezentující konkrétní řešení (tedy jedince). Při tvorbě nové generace dochází ke kombinování genetických informací rodičů. Způsob, jakým vzniká nový jedinec (jeho chromozom), je určen genetickými operátory selekce, křížení a mutace. Ty určují výběr rodičů, způsob kombinace jejich genetické informace a její následnou modifikaci případnou mutací.

Klíčovým pojmem genetických algoritmů je *fitness funkce*, která určuje kvalitu daného řešení

vzhledem k řešené úloze. Každému jedinci je přiřazena číselná hodnota vyjadřující jeho zdatnost, která ovlivňuje pravděpodobnost jeho výběru pro reprodukci.

Algoritmus pracuje s populací jedinců, která se v průběhu generací vyvíjí. V každé generaci jsou preferováni jedinci s vyšší hodnotou fitness, kteří mají větší pravděpodobnost předat své vlastnosti dalším generacím (Koza, 1994).

2.1 Schéma algoritmů

Zjednodušeně můžeme princip genetického algoritmu popsat následujícím pseudokódem:

```
inicializuj populaci P
vyhodnoť fitness každého jedince v P

while (není splněna podmínka ukončení):
    vytvoř novou populaci P'

    zachovej nejlepší jedince (elitismus)

    while (|P'| < velikost populace):
        vyber rodiče p1, p2 (selekce)
        vytvoř potomka (křížení)
        aplikuj mutaci
        přidej potomka do P'

    nahraď P ← P'
    vyhodnoť fitness

vrať nejlepší nalezené řešení
```

Jako podmínku ukončení můžem zvolit dosažení maximálního počtu generací nebo nalezení dostatečně kvalitního řešení (Lingaraj, 2016; Trinh, 2025).

2.2 Reprezentace řešení

Způsob reprezentace jedince závisí na typu řešeného problému, například:

- binární reprezentace, kde je řešení kódováno jako řetězec bitů, což je klasický přístup využívaný v raných genetických algoritmech,
- pro spojité optimalizační úlohy jsou jedinci reprezentováni jako vektory reálných čísel,
- pro kombinatorické úlohy, jako je problém obchodního cestujícího, se používá permutační reprezentace, kde chromozom představuje pořadí navštívených prvků.

Volba vhodné reprezentace má zásadní vliv na efektivitu genetického algoritmu, protože ovlivňuje návrh genetických operátorů i schopnost algoritmu prohledávat stavový prostor (Le a spol., 2023).

2.3 Genetické operátory

Prohledávání stavového prostoru genetickými algoritmy probíhá pomocí genetických operátorů - selekce, křížení a mutace.

2.3.1 Operátor selekce

Operátor selekce určuje, kteří jedinci budou vybráni jako rodiče pro vytvoření nové generace. Cílem je zvýhodnit kvalitnější jedince, tedy ty s co nejlepší fitness, aniž by došlo ke ztrátě diverzity populace.

V této práci byla použita *turnajová selekce*, při které je náhodně vybrána skupina jedinců a z ní je zvolen nejlepší, přičemž se proces opakuje pro výběr každého rodiče. Výhodou této metody je její jednoduchost a možnost řídit selekční tlak pomocí velikosti turnaje (Miller a Goldberg, 1995).

Další často používanou metodou je *ruletová selekce* (*Roulette Wheel Selection*), kde je pravděpodobnost výběru jedince úměrná jeho fitness, respektive poměru jeho fitness vůči celkovému součtu fitness všech jedinců. Alternativou je také *rank-based selekce*, která přiděluje pravděpodobnosti na základě pořadí jedinců, čímž omezuje vliv extrémních hodnot fitness (Lingaraj, 2016; BAJPAI a Kumar, 2010).

2.3.2 Operátor křížení

Křížení (*crossover*) slouží ke kombinaci genetické informace dvou rodičů za účelem vytvoření potomka. Tento operátor umožňuje prozkoumávat nové oblasti stavového prostoru.

Pro spojité úlohy byl použit *aritmetický crossover*, při kterém je potomek vytvořen jako lineární kombinace rodičů:

$$potomek = \alpha p_1 + (1 - \alpha) p_2 \tag{1}$$

Pro problém obchodního cestujícího byl použit operátor *Order Crossover* (OX), který zachovává relativní pořadí prvků a produkuje validní permutace.

Mezi další běžně používané operátory patří například jednobodové nebo dvoubodové křížení, typické pro binární reprezentaci, nebo *uniformní crossover*, který kombinuje geny rodičů na základě náhodné masky. U permutačních úloh se často využívají operátory jako PMX (*Partially Mapped Crossover*), které zachovávají validitu permutace (Lingaraj, 2016).

2.3.3 Operátor mutace

Mutace slouží k zavedení náhodných změn do jedinců a zajišťuje diverzitu populace. Pomáhá algoritmu uniknout z lokálních optim.

Pro spojité úlohy byla mutace realizována přidáním gaussovského šumu k jednotlivým proměnným. U problému obchodního cestujícího

byla použita *swap mutace*, při které dojde k prohození dvou náhodně vybraných prvků v permutaci.

Mezi další často používané mutační operátory patří například *uniformní mutace*, která nahrazuje hodnotu genu náhodně zvolenou hodnotou z definičního oboru, nebo *neuniformní mutace*, jejíž velikost změny se v průběhu evoluce postupně snižuje. U permutačních úloh se kromě swap mutace využívá také *inverzní mutace*, která obrací pořadí prvků v části chromozomu (Lingaraj, 2016).

2.4 Parametry GA

Výkonnost genetického algoritmu závisí na několika základních parametrech:

- velikost populace,
- pravděpodobnost křížení P_c,
- pravděpodobnost mutace P_m,
- počet generací.

Tyto parametry ovlivňují rovnováhu mezi explorací (prohledávání nových oblastí) a exploitací (vylepšování existujících řešení) (Trinh, 2025).

3 Nastavení experimentů

Za cíl našich experimentů jsme si stanovili analýzu vlivu nastavení parametrů genetického algoritmu na kvalitu nalezeného řešení, rychlost konvergence a stabilitu výsledků, přičemž jsme se zabývali hlavně porovnáním vlivu pravděpodobnosti mutace P_m. Mimo zkoumání vlivu parametrů jsme porovnávali také přístup s fixními parametry a adaptivním řízením pravděpodobnosti mutace.

3.1 Testovací úlohy

Pro experimentální analýzu jsme zvolili tři typy optimalizačních úloh, reprezentující různé charakteristiky stavového prostoru a různou obtížnost.

3.1.1 Jednorozměrná optimalizační úloha

Jako jednorozměrnou optimalizační úlohu jsme zvolili funkci:

$$y = sin(x) \cdot x \tag{2}$$

s definičním oborem $\langle 0; 20 \rangle$. Tato úloha slouží jako základní testovací scénář s relativně jednoduchou strukturou fitness funkce.

3.1.2 Dvourozměrná optimalizační úloha

Dvourozměrná úloha byla definována funkcí:

$$y = sin(x) \cdot cos(y) + 0,1x + 0,1y \tag{3}$$

na definičním oboru $(x, y) \in \langle 0; 20 \rangle \times \langle 0; 20 \rangle$.

Tato funkce kombinuje periodickou složku s lineárním trendem, čímž vytváří komplexní fitness krajinu s více lokálními extrémy. Úloha je vhodná pro analýzu schopnosti algoritmu překonávat lokální optima.

3.1.3 Problém obchodního cestujícího (TSP)

Třetí testovací úlohou byl problém obchodního cestujícího. TSP je klasický optimalizační problém, jehož cílem je najít nejkratší možnou cestu, která prochází všemi zadanými městy (body), přičemž každé město navštíví právě jednou a vrátí se do výchozího bodu. Jedná se o tzv. NP-težký problém, což znamená, že s počtem měst roste počet možných kombinací exponenciálně. Pro pouhých $n = 10$ měst je počet možných kombinací přes $18\,000$. Řešení této úlohy hrubou silou pro vetší počet měst tedy nepřipadá v úváhu, díky čemuž je vhodná pro testování genetických algoritmů (Cook a spol., 1997).

Bylo uvažováno $n = 30$ měst, což je přibližně $4,42 \times 10^{30}$, jejichž souřadnice byly generovány náhodně v rovině. Pro zajištění reprodukovatelnosti byly souřadnice měst generovány pouze jednou a následně použity ve všech bězích experimentu (Matai a spol., 2010).

Fitness funkce byla definována jako převrácená hodnota celkové délky uzavřené cesty d:

$$f = \frac{1}{d} \tag{4}$$

3.2 Nastavení genetických algoritmů

Pro všechny experimenty byla použita následující konfigurace genetického algoritmu:

- velikost populace: 80 a 180 jedinců,
- počet generací: 30 pro 1D funkci, 150 pro 2D funkci, 400 pro TSP,
- pravděpodobnost křížení: $P_c = 0,8$ a $P_c = 0,5$,
- pravděpodobnost mutace: $P_m = 0,01$, $P_m = 0,05$ a $P_m = 0,1$
- selekce: turnajová selekce (velikost turnaje 3),
- elitismus: nejlepší jedinec byl vždy zachován do další generace.

Pro spojité úlohy byla mutace realizována přidáním gaussovského šumu, zatímco u TSP byla použita swap mutace. Pro křížení byl u spojitých úloh použit aritmetický operátor, zatímco u TSP byl aplikován operátor Order Crossover (OX).

Data pro každou konfiguraci algoritmu byla vyhodnocována na základě 30 běhů.

3.3 Adaptivní řízení parametrů

V adaptivní variantě genetického algoritmu byla dynamicky upravována pravděpodobnost mutace P_m v závislosti na průběhu optimalizace. Cílem bylo reagovat na stagnaci algoritmu a současně podpořit jemnější doladění řešení v případě úspěšného vývoje.

Pokud v aktuální generaci došlo ke zlepšení nejlepší dosažené hodnoty fitness oproti dosud nalezenému maximu, byla pravděpodobnost mutace snížena vynásobením koeficientem 0,9. Tím docházelo k postupnému omezení náhodnosti a podpoře exploatace okolí kvalitního řešení.

Naopak pokud během 10 po sobě jdoucích generací nedošlo ke zlepšení hodnoty fitness (stav stagnace), byla pravděpodobnost mutace zvýšena. U jednorozměrné a dvourozměrné optimalizační úlohy byl použit koeficient 1,5, který umožňuje rychlé zvýšení diverzity populace a intenzivnější exploraci nových oblastí stavového prostoru.

U problému obchodního cestujícího byl zvolen nižší koeficient 1,1. Důvodem je diskrétní charakter problému a delší evoluce, kde příliš agresivní navyšování mutace může vést k narušení již nalezených kvalitních částí řešení a ke ztrátě konvergence. Jemnější adaptace proto umožňuje postupnější obnovu diverzity bez výrazného narušení dosavadního vývoje populace.

Pravděpodobnost mutace byla během celého běhu omezena předem definovanými dolními a horními mezemi, aby nedocházelo k její přílišné redukci nebo naopak k nadměrně náhodnému chování algoritmu.

4 Výsledky experimentů

V této kapitole zhodnotíme data získaná z experimentů. Pro každou úlohu zde uvádíme tabulku se základní statistikou a čtyři grafy, které hodnotí:

- rychlost konvergence pro různá P_m,
- rychlost konvergence pro stejné P_m a jiné P_c a velikost populace (pop_size),
- variabilitu generace, ve které bylo nalezeno nejlepší řešení ($best_gen$) pro různá P_m,
- variabilitu $best_gen$ pro stejné P_m a jiné P_c a pop_size.

Pro zkrácení zápisu jednotlivých algoritmů jsme zavedli následující značení: Problém-varianta-pop_size x Pc x Pm. Označení GA, který řeší 1D úlohu, má fixně nastavené parametry, $pop_size = 80$, $P_c = 0,8$ a $P_m = 0,01$ tak bude vypadat takto: $1Dfix80x0,8x0,01$.

4.1 Jednorozměrná optimalizační úloha

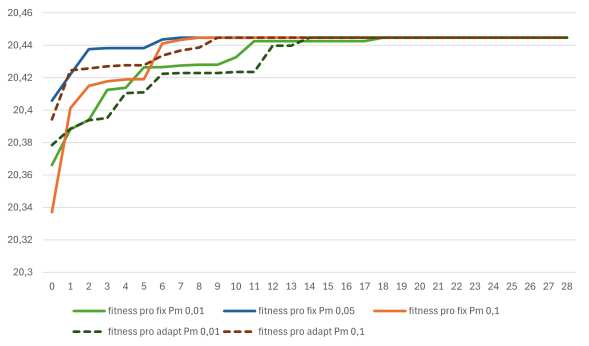
Tato podkapitola se zaměřuje na vyhodnocení genetického algoritmu při řešení jednorozměrné optimalizační úlohy. Porovnávány jsou jak varianty s fixními parametry, tak adaptivní přístup. Výsledky jsou shrnuty v Tabulce 1 a doplněny grafy průběhu konvergence a variability.

Tab. 1: Porovnání výsledků různých nastavení parametrů GA pro 1D úlohu.

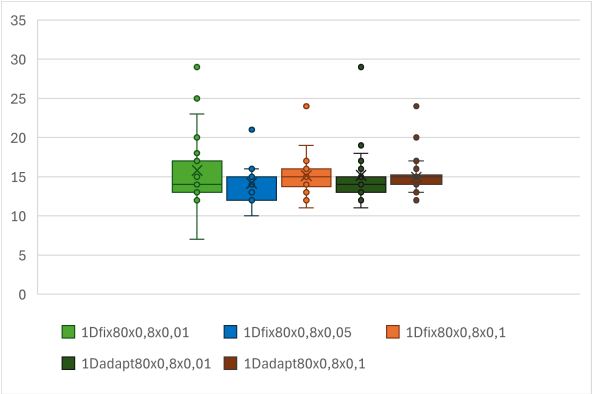
Alg.	Mean fit	Std fit	Med gen	Std gen
fix80x0,8x0,01	20,4447831	5,1E-06	14	4,88
fix80x0,8x0,05	20,4447826	7,9E-06	15	2,02
fix80x0,8x0,1	20,4447841	7,2E-15	15	2,42
adapt80x0,8x0,01	20,4447841	7,2E-15	14	4,12
adapt80x0,8x0,1	20,4447841	7,2E-15	15	2,37
fix180x0,8x0,01	20,4447841	7,2E-15	14	1,19
fix80x0,5x0,01	20,4447840	2,8E-07	15	3,02

Výsledky ukazují, že všechny testované konfigurace genetického algoritmu dosahují prakticky identické hodnoty fitness ($\approx 20,44478$), což odpovídá globálnímu maximu dané funkce. Rozdíly mezi jednotlivými nastaveními se proto projevují především v rychlosti konvergence a stabilitě, i když jen nevýrazně v porovnání se složitějšími úlohami.

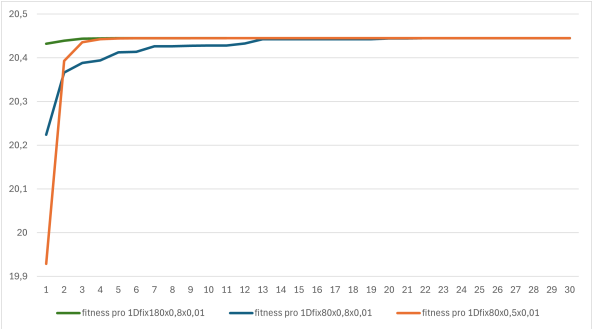
Z Tabulky 1 a Obrázku 1 můžeme vyčíst, ž vliv mutace je v této úloze minimální, což je dáno jednoduchostí optimalizačního problému. Ani adaptivní řízení parametrů zde nepřináší významné zlepšení oproti fixnímu nastavení.



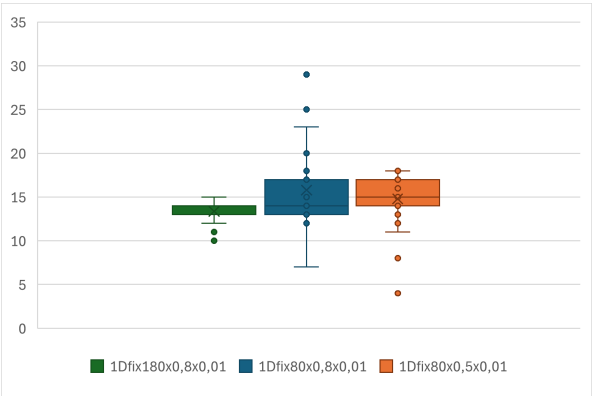
Obr. 1: Porovnání rychlosti konvergence a kvality řešení pro GA 1Dfix80x0,8xPm a 1Dadapt80x0,8xPm.



Obr. 2: Porovnání variance *best_gen* pro GA 1Dfix80x0,8xPm a 1Dadapt80x0,8xPm.



Obr. 3: Porovnání rychlosti konvergence a kvality řešení pro GA $P_m = 0,01$.



Obr. 4: Porovnání variance *best_gen* pro GA $P_m = 0,01$.

Z hlediska konvergence dosahují nejlepší výsledky konfigurace s vyšší populací (1Dfix180x0,8x0,01), která vykazuje nejnižší variabilitu (*Std. best_gen = 1,19*) při zachování nízkého mediánu generace konvergence. Naopak nižší pravděpodobnost křížení (1Dfix80x0,5x0,01) vede k pomalejší a méně stabilní konvergenci, jak je vidět na Obrázcích 3 a 4.

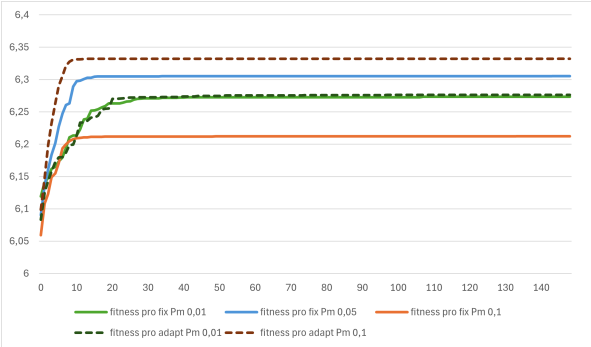
4.2 Dvourozměrná optimalizační úloha

V této podkapitole jsou prezentovány výsledky pro dvourozměrnou optimalizační úlohu, která představuje složitější fitness krajinu s větším počtem lokálních extrémů. Přehled dosažených výsledků je uveden v Tabulce 2 a doplněn grafy.

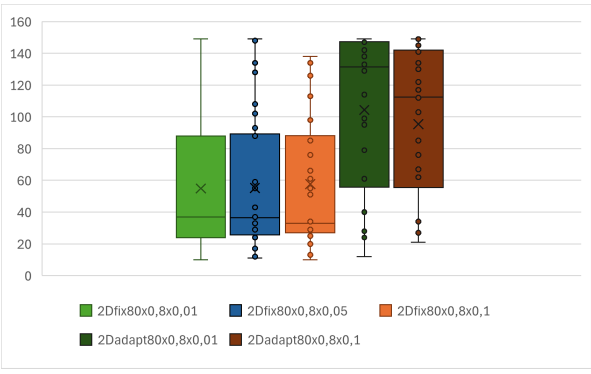
Tab. 2: Porovnání výsledků různých nastavení parametrů GA pro 2D úlohu.

Alg.	Mean fit	Std fit	Med gen	Std gen
fix80x0,8x0,01	6,2733	0,3260	37	42,41
fix80x0,8x0,05	6,3052	0,2894	36,5	42,32
fix80x0,8x0,1	6,2121	0,2331	33	42,08
adapt80x0,8x0,01	6,2765	0,2375	131,5	49,89
adapt80x0,8x0,1	6,3323	0,2876	112,5	45,62
fix180x0,8x0,01	6,4568	0,2523	30	27,78
fix80x0,5x0,01	6,2491	0,2332	36,5	38,02

Na rozdíl od 1D úlohy se zde již výrazněji projevují rozdíly mezi jednotlivými konfiguracemi. Jak je vidět z Tabulky 2 a na Obrázku 7, nejlepší hodnoty fitness dosahuje konfigurace s větší populací (2Dfix180x0,8x0,01), což potvrzuje význam diverzity populace u složitějších problémů.



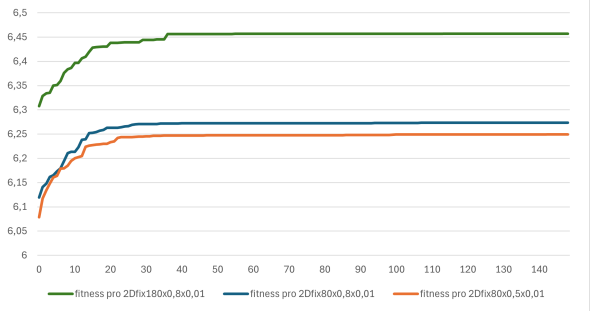
Obr. 5: Porovnání rychlosti konvergence a kvality řešení pro GA 2Dfix80x0,8xPm a 1Dadapt80x0,8xPm.



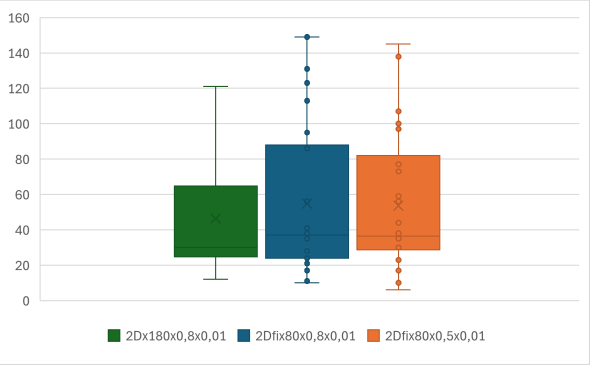
Obr. 6: Porovnání variance *best_gen* pro GA 2Dfix80x0,8xPm a 2Dadapt80x0,8xPm.

Vyšší pravděpodobnost mutace ($P_m = 0,1$) vede ke zhoršení kvality řešení, což naznačuje přílišnou míru náhodnosti. Naopak střední hodnota mutace

($P_m = 0,05$) poskytuje vyvážený kompromis mezi explorací a exploitací. Tato konfigurace dosáhla také dobré stability, jak můžeme názorně vidět na Obrázku 6



Obr. 7: Porovnání rychlosti konvergence a kvality řešení pro GA $P_m = 0,01$.



Obr. 8: Porovnání variance *best_gen* pro GA $P_m = 0,01$.

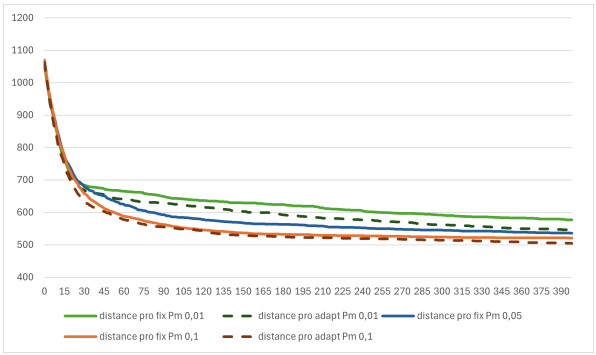
Adaptivní genetický algoritmus dosahuje srovnatelných nebo mírně lepších hodnot fitness (např. 2Dadapt80x0,8x0,1), avšak za cenu výrazně pomalejší konvergence (medián až 131,5 generace). Tento výsledek ukazuje, že adaptivní řízení podporuje důkladnější prohledávání prostoru, ale může zpomalit nalezení řešení.

4.3 Problém obchodního cestujícího

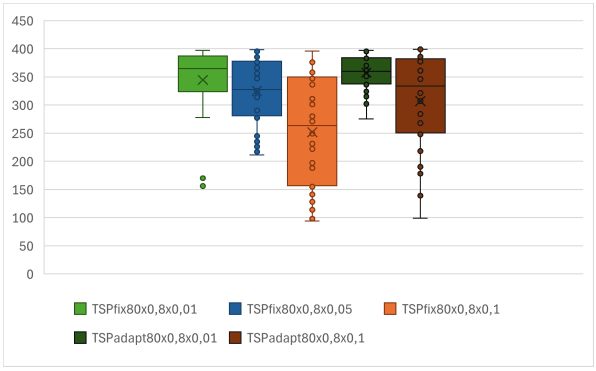
Tato podkapitola se zabývá aplikací genetického algoritmu na problém obchodního cestujícího, který představuje diskrétní kombinatorickou úlohu. Výsledky jsou shrnuty v Tabulce 3 a opět doplněny grafy. Pro lepší interpretaci tentokrát místo hodnoty fitness funkce používáme jako ukazatel kvality nalezeného řešení celkovou délku cesty *distance*.

Tab. 3: Porovnání výsledků různých nastavení parametrů GA pro TSP.

Alg.	Mean dis	Std dis	Med gen	Std gen
fix80x0,8x0,01	576,69	45,47	365	59,33
fix80x0,8x0,05	535,90	38,49	327,5	57,35
fix80x0,8x0,1	521,00	28,23	263,5	99,01
adapt80x0,8x0,01	546,16	47,10	360	31,02
adapt80x0,8x0,1	504,60	27,92	333,5	85,12
fix180x0,8x0,01	510,06	36,55	322	93,08
fix80x0,5x0,01	613,19	40,44	349	44,88



Obr. 9: Porovnání rychlosti konvergence a kvality řešení pro GA různé P_m.



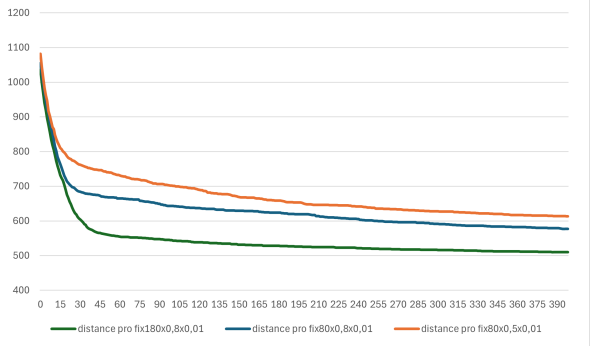
Obr. 10: Porovnání variance *best_gen* pro GA TSPfix80x0,8xPm a TSPadapt80x0,8xPm.

U problému obchodního cestujícího jsou rozdíly mezi konfiguracemi nejvýraznější. Nejlepších výsledků dosahuje adaptivní genetický algoritmus (TSPadapt80x0,8x0,1), který minimalizuje délku trasy na průměrnou hodnotu 504,60, jak je vidět v Tabulce 3 a Obrázku 9. Zajímavé je, že TSPadapt80x0,8x0,1 oproti TSPfix80x0,8x0,1 vykazuje menší zlepšení, než TSPadapt80x0,8x0,01 proti TSPfix80x0,8x0,01. U 2D úlohy jsme přitom mohli pozorovat opačný jev.

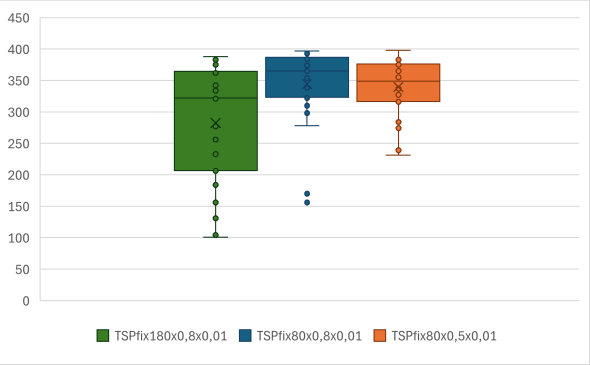
Vyšší mutace ($P_m = 0,1$) se zde ukazuje jako výhodná, protože pomáhá překonávat lokální optima. Naopak nízká mutace vede k horším výsledkům (např. právě TSPfix80x0,8x0,01).

Zvýšení velikosti populace (TSPfix180x0,8x0,01) rovněž vede ke zlepšení kvality řešení, avšak za cenu vyšší variability konvergence. Snížení pravděpodobnosti křížení (TSPfix80x0,5x0,01) má ne-

gativní dopad na kvalitu řešení i rychlost konvergence, jak je možné vyčíst z Tabulky 3 a Obrázků 11 a 12.



Obr. 11: Porovnání rychlosti konvergence a kvality řešení pro GA $P_m = 0,01$.



Obr. 12: Porovnání variance *best_gen* pro GA $P_m = 0,01$.

Adaptivní přístup se v této úloze ukazuje jako nejefektivnější, protože umožňuje dynamicky reagovat na stagnaci a lépe vyvažovat exploraci a exploitaci.

5 Závěr

Výsledky experimentů potvrzují, že vliv nastavení parametrů genetického algoritmu je silně závislý na složitosti řešené úlohy. Zatímco u jednoduchých optimalizačních problémů je možné dosáhnout kvalitních výsledků i s pevně nastavenými parametry, u složitějších úloh se význam vhodného nastavení výrazně zvyšuje.

Pozorovaný kompromis mezi rychlostí konvergence a kvalitou výsledného řešení odpovídá známému konsenzu mezi explorací a exploitací, který je v literatuře považován za jeden z klíčových principů adaptivního řízení parametrů (Eiben a Smith, 2003; Eiben a spol., 2007).

V případě jednorozměrné úlohy dosahovaly všechny testované konfigurace prakticky totožné hodnoty fitness odpovídající globálnímu maximu. Rozdíly mezi jednotlivými nastaveními se projevily především v rychlosti konvergence a její stabilitě, i když ne příliš výrazně. Vliv pravděpodobnosti mutace byl minimální a adaptivní řízení parametrů nepřineslo významné zlepšení. To odpovídá charakteru úlohy s jednoduchou fitness krajinou bez výrazných lokálních extrémů.

U dvourozměrné úlohy se již projevily rozdíly mezi jednotlivými konfiguracemi. Bylo pozorováno, že střední hodnota mutace představuje vhodný kompromis mezi explorací a exploitací, zatímco příliš vysoká mutace vede k degradaci kvality řešení. Adaptivní přístup zde umožnil dosahovat srovnatelných nebo mírně lepších výsledků, avšak za cenu pomalejší konvergence. Tento jev lze interpretovat jako důsledek intenzivnějšího prohledávání stavového prostoru.

Nejvýraznější rozdíly byly pozorovány u problému obchodního cestujícího. V tomto případě se adaptivní genetický algoritmus ukázal jako nejefektivnější přístup, který dosahoval nejlepších hodnot cílové funkce. Vyšší pravděpodobnost mutace se ukázala jako klíčová pro překonávání lokálních optim, zatímco nízká mutace vedla k předčasné konvergenci. Adaptivní řízení umožnilo dynamicky reagovat na stagnaci algoritmu a tím zlepšit kvalitu výsledků.

Napříč všemi úlohami se také potvrdil vliv velikosti populace. Větší populace obecně vedla k lepším výsledkům a stabilnější konvergenci, avšak za cenu vyšší výpočetní náročnosti. Naopak snížení pravděpodobnosti křížení mělo spíše negativní dopad, zejména u složitějších úloh.

Zajímavým zjištěním byl rozdílný vliv adaptivního řízení parametrů v závislosti na počáteční hodnotě mutace. U problému obchodního cestujícího byl přínos adaptivního přístupu výraznější při nízké počáteční hodnotě mutace, zatímco při vyšší mutaci byl rozdíl oproti fixnímu nastavení menší. Tento jev lze vysvětlit tím, že nízká mutace vede k předčasné konvergenci, kterou adaptivní mechanismus dokáže překonat zvýšením diverzity populace.

Naopak u spojité dvourozměrné úlohy byl pozorován opačný trend, kdy adaptivní přístup přinášel větší zlepšení při vyšší počáteční hodnotě mutace. V tomto případě adaptivní mechanismus umožňuje postupné snižování mutace v okolí maxima funkce a tím přejít od ”prohledávání”k ”ladění”. Výsledky tak ukazují, že efektivita adaptivního řízení parametrů je úzce spojena s charakterem řešené úlohy.

Celkově lze konstatovat, že adaptivní řízení parametrů je nejpřínosnější u komplexních optimalizačních problémů, kde pomáhá udržovat rovnováhu mezi explorací a exploitací. U jednodušších úloh však jeho přínos není významný a fixní nastavení parametrů je dostatečné a výpočetně úspornější. Tento závěr je v souladu s dřívějšími přehledovými pracemi, které uvádějí, že adaptivní mechanismy přinášejí největší výhody zejména u rozsáhlých a multimodálních optimalizačních úloh (Eiben a Smith, 2003; Eiben a spol., 2007)

Budoucí práce by se mohla zaměřit na komplexnější analýzu parametrů, analýzu citlivosti samotného adaptačního mechanismu (například vlivu velikosti kroku při změně pravděpodobnosti mutace), případně na vliv zvolených genetických operátorů. Zajímavým směrem je také testování na širší škále problémů a použití statistických testů pro detailnější vyhodnocení výsledků.

Reference

BAJPAI, P. a Kumar, M. (2010). Genetic algorithm - an approach to solve global optimization problems. *Indian Journal of Computer Science and Engineering*, 1:199–206.

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., Schrijver, A. a Sayer, P. (1997). *Combinatorial optimization*. John Wiley Sons Incorporated.

Eiben, A., Michalewicz, Z., Schoenauer, M. a Smith, J. (2007). Parameter Control in Evolutionary Algorithms. Lobo, F. G., Lima, C. F. a Michalewicz, Z. (eds.), V *Parameter Setting in Evolutionary Algorithms*, vol. 54 z *Studies in Computational Intelligence*, s. 19–46. Springer Verlag.

Eiben, A. E. a Smith, J. E. (2003). *Introduction to Evolutionary Computing*. Springer.

Koza, J. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, 4(2).

Le, K., Wei, T. a Tagalogon, J. (2023). Genetic algorithm and application. *Genetic algorithm and Application*.

Lingaraj, H. (2016). A study on genetic algorithm and its applications. *International Journal of Computer Sciences and Engineering*, 4:139–143.

Matai, R., Singh, S. a Mittal, M. (2010). *Traveling Salesman Problem: an Overview of Applications, Formulations, and Solution Approaches*.

Miller, B. a Goldberg, D. E. (1995). *Genetic algorithms, tournament selection and the effects of noise*.

Trinh, M. V. T. K. (2025). Research on genetic algorithms: Concepts, models, and applications. *International Journal of Education and Social Science Research (IJESSR)*, 8:354–373.

Does a transformer think like a human?

Andrej Lúčny

Katedra aplikovanej informatiky, Fakulta matematiky, fyziky a informatiky, Univerzita Komenského
KAI FMFI UK, Mlynská dolina, 842 48 Bratislava
lucny@fmph.uniba.sk

Abstract

We describe the operation of transformers and compare it with our ideas about the nature of human information processing. Using examples of the use of individual models, we point out striking similarities and fundamental differences between a transformer and a human:
1.) By setting the correct temperature, chatbots can generate logical errors characteristic of humans (we proposed these years ago as a test of whether an artificial intelligence system resembles a human).
2.) Using the example of holistic OCR, we show that it is possible to loop and confuse a transformer in a way that a human would not make a mistake.
3.) In the case of naming a controversial image, we demonstrate behavior that is not typical of humans and overcome it using an agnostic algorithm for calculating bipartition, suggesting that humans probably also have a similar ability.

1 Introduction

Alan Turing had shown in Turing (1950) that an essential property of artificial intelligence systems is that they make mistakes. His argument was based on Gödel's incompleteness theorems, from which it is possible to derive that for any mistake-free intelligent system, there is an input on which it endlessly loops, Kelemen (1990). In this paper, we relate these old ideas to the latest transformer-based artificial intelligence systems: ChatGPT - OpenAI (2023), DeepSeekOCR - Wei (2025), and CLIP – Radford (2021). We find interesting relations.

2 How transformer-based LLMs works

A large language model is a neural network with a transformer architecture that, given a text as a sequence of tokens (pieces), calculates for each token the probability that it follows. If the transformer is of the decoder-only type (for example, GPT, the basis of the most famous chatbot), we use it as a generator of a scenario that starts with a user question and continues with the assistant's answer. From the entered text, it calculates the probabilities, based on which it selects the following token. That token is appended to the current scenario and then fed back into the transformer, run again to generate the next token. It repeats this until it generates a special token representing the end of the sequence (EOS) (see Fig. 1).

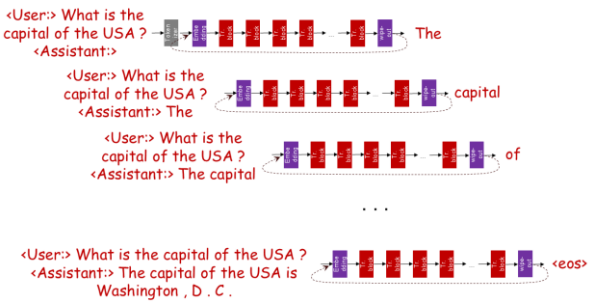


Fig. 1. Decoder-only transformer in the action

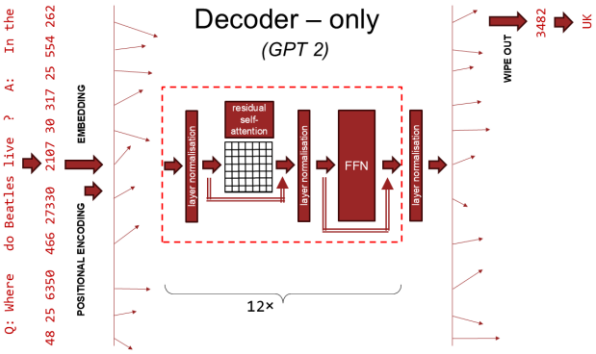


Fig. 2. The next-token prediction (GPT2)

The next-token calculation process (Fig. 2) begins with embedding, which assigns each token a vector in a multidimensional space. The dimension of this space corresponds to the number of features by which two things can differ in their meaning. Since there are only a few tens of thousands of possible tokens, embedding is implemented using a table that maps token indices to their corresponding vectors. The table values are the network parameters; their final form is the result of training. The vectors are organized in the space so that the angle between them is small when they occur frequently in the same context in the training data (and therefore affect each other's meaning). The almost equal lengths of the vectors are ensured by layer normalization, a building block of the transformer,

which is applied after each data processing step. As a result of its application, the vectors after each normalization roughly point to the surface of a hypersphere with a dimensionality one less than that of the vectors.

Once the input text has been transformed into a list of vectors, the so-called hidden states, we add a positional code to each vector that encodes its position in the sentence. We further transform them into other hidden states by a sequence of so-called layers. The number of hidden states in each layer is the same and corresponds to the size of the input. Each layer consists of two modules: a residual multi-headed self-attention and a perceptron.

The first module, self-attention, refines and contextualizes the meaning of hidden states. It works by mixing each hidden state with a little of the other hidden states, mainly those that yield the largest scalar products, i.e., the smallest angles. Attention includes a scaling coefficient that determines how much of the more similar and how much of the more different hidden states are mixed. Its optimal value is the square root of the hidden-states dimension, but a slight change does not cause a system crash; it only leads to slightly different behavior. To get slightly different outputs from the transformer when different calls are made with the same input, we multiply it by the number $(1+T)$, where T is the so-called temperature, a randomly generated non-negative number. The higher the temperature, the more "creative" the generated statement is, until it becomes meaningless. Therefore, its reasonable maximum is considered to be the value of two.

The second module processes each hidden state separately with a perceptron. In doing so, it transforms the features of the hidden states into others, allowing nonlinear states to emerge between the new and original features.

After processing the hidden states across all layers, we wipe out the last state to index the next token. Unlike inference, during training we enforce wipeout across all hidden states, requiring that the next one emerges from the last, the last from the next-to-the-last, and so on.

3 As good as humans

Many years ago, in Lúčny (2002), we proposed a specific test to determine whether an artificial system is intelligent in a manner characteristic of humans. The tests consisted of whether it was capable of failing in the same way as a human. One such test consisted of answering two questions:

1. How many dollars are in a dozen?
2. And how many quarters?

Only a few people answer this question with 12 and 12. Most people answer 12 and 48. But there are also more exotic answers, such as 12 and 3. It may be true to object that the question is not quite correctly understood as "How many quarters are in a dozen dollars?" But people do not stop being wrong even if we insist:

1. How many dollars are in a dozen?
2. And how many quarters are in a dozen?

Is the transformer wrong like this? It's shocking, but it's wrong exactly like this when we increase the temperature to get various answers (Fig 3, Fig. 4). We even understand why quite clearly. The token "dollar" is simply contextually stronger for "quarter" than the rarely occurring "dozen", which, moreover, denotes quantity and not currency. We can measure in the latent space that "dollar" and "dozen" make an angle of 68°. And "dollar" and "dozen" make 67°, while "quarter" and "dozen" make 72°. Moreover, if we try to define the terms in the system prompt, it almost always fails, as humans do (Fig. 5).

The transformer has its limits as a means of logical reasoning. And so do we. The transformer is not as strongly intelligent as we are. We are as weakly intelligent as the transformer.

```
client = OpenAI(api_key=api_key)

SYSTEM_PROMPT = (
  #"A dollar consists of four quarters. "
  #"A dozen refers to a group or set of twelve items. "
  "Answer with a single word. "
)
QUESTION_1 = "How many dollars are in a dozen?"
ANSWER_1 = "Twelve."
#QUESTION_2 = "And how many quarters?"
QUESTION_2 = "And how many quarters are in a dozen?"

temperature = 0.0 # 0.0 .. 2.0
messages = [
  {"role": "system", "content": SYSTEM_PROMPT},
  {"role": "user", "content": QUESTION_1},
  {"role": "assistant", "content": ANSWER_1},
  {"role": "user", "content": QUESTION_2},
]

response = client.chat.completions.create(
  model="gpt-4",
  messages=messages,
  temperature=temperature,
  max_tokens=50,
)

assistant_text = response.choices[0].message.content.strip()
print(assistant_text) # Forty-eight.
```

Fig. 3. A code snippet for calling GPT4o

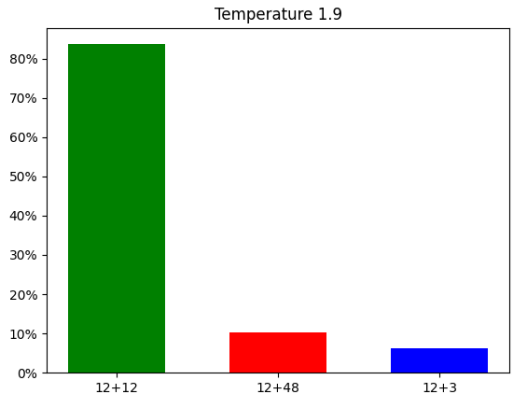


Fig. 4. Results in the human-like intelligence test

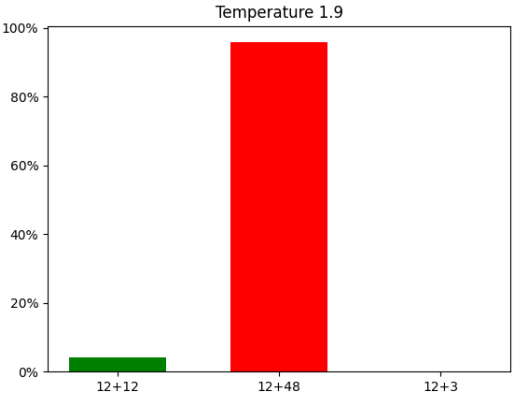


Fig. 5. Results when we explicitly define terms

4 Much worse than humas

On the other hand, we can also get behavior from the transformer that humans would never produce. In 2025, DeepSeekOCR broke all records in OCR benchmarks. It correctly reads most texts. However, it is completely unusable in industrial practice. Why?

DeepSeekOCR is an encoder-decoder transformer (Fig. 6). More precisely, it has two encoders: a visual encoder and a text encoder. Both produce hidden states that are combined and, through the so-called cross-attention, influence the generation of a text response. It is generated gradually, with each pass through the decoder, we receive only one more token of recognized text. We repeat this until we get EOS. And unfortunately, it is not that difficult to find an image we will never get EOS. It is enough to present the model with sufficiently hatched text (Fig. 7). Probably the hatching induces the presence of many ones, and having read many ones, the model expects to read two, three, and so on; the prediction coming from the read overwhelms the observation coming from the seen.

But even when DeepSeekOCR responds with a twist, it can still produce cute errors, stemming from the fact that the hidden states of the image and the text prompt are coupled, meaning their representations are similar across modalities. So when you prompt it: “read this text three times,” it can do it. But when you prompt it: “read this text,” and the image contains the text “read this text three times,” it sometimes follows the instruction from the image.

So this model does exactly what we fear most from artificial intelligence: that after a long period of flawless functioning, it suddenly fails fatally. But not only that. It does exactly what we consider inhuman in an artificial system.

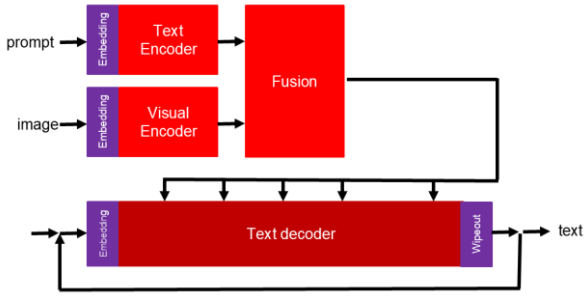


Fig. 6. Schema of transformer-based OCR



Fig. 7. Perfect reading versus endless looping via DeepSeekOCR with the default prompt: `"<image>\n<|grounding|>Convert the document to markdown. "`

5 Agnostic capabilities

CLIP is a dual-encoder-only transformer consisting of separate text and visual encoders. It is trained on many images with text captions, so that text and image representations share the same latent space. As a result, it is possible to turn it into an image classifier of categories described by textual descriptions. When we get textual labels of the COCO dataset categories (containing “bird” and “dog”) and encode them with the CLIP text encoder, we can classify images by the maximal cosine similarity of these representations with

the representation obtained by encoding an image by the CLIP visual encoder.

CLIP works quite well until we present it with an image, as in Fig. 8 (left). Then its response is very far from human-like. It claims we see a dog while we clearly see a bird, labeled as a dog. Humans clearly recognize that the image consists of two controversial parts, whereas CLIP lacks analogical ability. However, it is possible to add it. CLIP is a visual transformer, so it internally processes images into feature maps. It splits the image into regions described by feature vectors. They share the same latent space, so we can try to partition them into two parts, such that vectors are rather similar in each part and rather different between the parts. We perform this bipartitioning in a class-agnostic manner using a suitable measure. In Lucny (2026), we designed a fast algorithm for bipartitioning based on the measure Normalized Cut proposed in Shi-Malik (2000). So we take the feature map calculated by the CLIP image encoder, and by the algorithm (FastNCut), we split it into two parts that we can name by CLIP separately. The result is impressive; we get a bird in one part and a dog in the other part (Fig. 8, right).

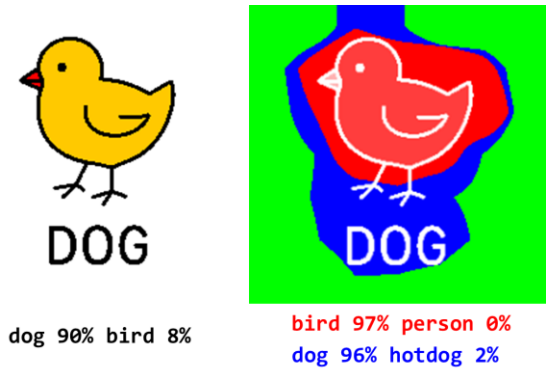


Fig. 8. An example of controversy named by CLIP and named partitioned by an agnostic algorithm FastNCut

6 Conclusion

We presented cases in which transformers exhibit both human-like and inhuman-like behavior. On the one hand, we find that the transformers are more human-like than any other machine. They provide similar quality and also similar mistakes. As we are not able to pick knowledge from our brains, the same is true for transformers. On the other hand, from the structural point of view, our brains and transformers are simply two different things. Transformers can also fail in a very inhuman-like way.

To identify what humans have that the transformers do not (so far), we have suggested the ability to class-agnostically partition the feature vector sets, demonstrating its potential on a specific example.

Acknowledgement

This research was supported by the Horizon Europe MSCA project TRAIL, no. 101072488 and the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00048

References

Cristina, S., & Saeed, M. (2022). Building transformer models with attention: Implementing a neural machine translator from scratch in Keras. *Machine Learning Mastery*.

Kelemen, J. (1990). Myslenie, počítač. *Bratislava: Spektrum*.

Lúčny, A. (2002). Modelovanie kognitívneho zlyhania multiagentovým systémom so stigmergickou komunikáciou. *In J. Kelemen & V. Kvasnička (Eds.), Kognice a umělý život II: Sborník 2. česko-slovenského semináře* (13.–16. května 2002, Mílovy) Slezská univerzita v Opavě

Lúčny, Andrej, A Fast Algorithm for Normalized Cut with Applications on Bipartitioning Feature Maps in Deep Learning. *Available at SSRN*: http://dx.doi.org/10.2139/ssrn.6482332

OpenAI (2023). GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Ilya Sutskever. (2021). Learning transferable visual models from natural language supervision. *In Proceedings of the International Conference on Machine Learning (ICML)* (pp. 8748–8763). https://arxiv.org/abs/2103.00020

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8),* 888–905. https://doi.org/10.1109/34.868688

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59(236),* 433–460. https://doi.org/10.1093/mind/LIX.236.433

Wei, H., Sun, Y., & Li, Y. (2025). DeepSeek-OCR: Contexts optical compression. *arXiv:2510.18234*. https://arxiv.org/abs/2510.18234

Falzifikační charakter sociální agence v užívání vizuálně-generativní AI

Jan Šmejkal

Jan Šmejkal: Ústav filozofie a religionistiky FFUK,
Náměstí Jana Palacha 2,
honzi3702@gmail.com

Abstrakt

Příspěvek se zabývá schopností umělých neuronových sítí vygenerovat unikátní obrázek. Na první pohled není jasné, jak a komu lze v kontextu tohoto procesu atribuovat agenci. Prvním argumentem příspěvku je tvrzení, že agence leží jak na straně lidského operátora, tak na straně technologie, která vystupuje jako specifický tvořivý nástroj schopný emulovat styl lidského umělce. Jen lidský článek vztahu je však schopen intendovat a proto musí být agence v konečném důsledku přičtena jemu. Povaha této agence je dále ztotožněna se zastírací snahou a do uměleckého kontextu převedena apropriací pojmu falzifikace a falzifikátu. Argumentace se opírá o agenční teorii antropologa Alfreda Gella. Autor příspěvku nelegitimizuje jakékoliv praktiky spojené s provozováním ani trénováním neuronových sítí, které využívají obrazový materiál bez souhlasu jeho tvůrců.

1 Obrázky z umělé inteligence

Systémy umělé inteligence zaměřené na produkci a generování statického vizuálního obsahu[1] jsou trénovány na fotografiích a obrazech lidských profesionálů a umělců, dokáží emulovat jejich styl[2] a vyvést v něm vizuální dílo na základě slovního zadání. Tento proces běžným jazykem nazýváme „nechat si vygenerovat obrázek umělou inteligencí“. Ač je funkcionalita těchto sítí rozličná a služby, které se skrz ně nabízí, různorodé, v našem příspěvku se budeme soustředit na jejich vrcholnou dovednost, tedy schopnost přivést v existenci do té doby neexistující statické vizuální dílo („vygenerovat obrázek“).

Tato funkcionalita má samozřejmě dopad na trh s námezdným uměním a podstatným způsobem mění rozložení ekonomicko-kulturních sil na obou stranách vazby „poptávající-umělec“. Naše zkoumání se však bude zaobírat teoretickou otázkou atribuce umělecké agence a její povahy v kontextu „generování obrázků“ za pomocí neuronových sítí.

K tomu budeme využívat terminologii agenčního systému antropologa Alfreda Gella, jehož užití krátce zdůvodníme. Tato antropologická teorie byla svým autorem využívána zejména pro objektovou analýzu lidového umění domorodých kmenů. Ač vizuálně-generativní AI (dále jen VGAI)[3] není domorodý artefakt, přeci jen se v kontextu tradičních uměleckých forem jedná o něco cizorodého, co nezapadá do klasického schématu ani situace vizuálního umění. Také se tato teorie vyznačuje poměrnou liberalitou, se kterou nakládá s pojmy umění, dílo, artefakt atd., jenž mohou být v kontextu jiných uměleckých teorií exkluzivní a vykazovat rysy mnoha podob centrismu.[4] Posledním důvodem je to, že na rozdíl např. od agenční teorie Bruna Latoura je Gellova teorie pragmatičtější[5] a snazší na pochopení.

Zbytek úvodu využijeme pro postižení dvou klíčových *technologických* aspektů VGAI, které jsou pro naši analýzu podstatné a na něž omezíme vysvětlení těchto technologických artefaktů.

A) Grafické generátory na bázi *diffusion* jsou produktem obrácení funkcionality neuronových sítí sloužících k rozpoznávání tvarů a objektů: pracují na základě opakovaného včítání nejpravděpodobnějšího vzoru do náhodné plochy bílého šumu. (Bergman, Stryker, 2026) Produkt VGAI je tedy v nepřímé vazbě k vizuálním datům,[6] na kterých byla původní rozpoznávající síť trénována a z nichž získává vzory, jež jsou reprodukovány. Tato nepřímá vazba má povahu napodobení. Pro náš problém je podstatné, že ačkoliv VGAI generuje unikátní (ve smyslu neopakované) artefakty, pracuje vždy s určitým množstvím předloh a „inspirací“, které byly využity při jejím trénování a které tvořili lidští řemeslníci. Oproti

[1] V tomto příspěvku se soustředíme pouze na problematiku technologií generujících “obrázky”, tedy statické dvourozměrné vizuální objekty. Pod tímto označením myslíme např. technologie Imagen, DALL-E, Midjourney, Craiyon apod.

[2] Důrazu na stylovou emulaci si všímá Yao (Yao, 2025, str. 2). Stylu zde rozumíme jako ustrnulé interpretaci vizuální reality při jejím převodu na uměleckou reprezentaci.

[3] Volíme novotvar “vizuálně-generativní” namísto zaběhnutého “generativní”, protože generace vizuálních dat je pouze jednou z možných generativních funkcí AI, viz (Corchado et al., 2023)

[4] Např. eurocentrismu, antropocentrismu nebo manocentrismu.

[5] To je míněno doslova, jedná se o teorii hlásící se k pragmatismu.

[6] Tedy k databázi, která se tak stává tzv. *konceptuální metaforou* skutečného uměleckého světa (de Mul, 2009, str. 101)

kreativitě tedy VGAI sází na *variaci existujícího*. Tato vlastnost síť činí uživatelsky předvídatelnější a lépe monetizovatelnou.

B) Neuronové sítě fungují na stochastickém (pravděpodobnostním) základu. (Bergman, Stryker, 2026) To znamená, že namísto toho, abychom vůči zadání A dostali vždy výsledek A' (jako v deterministickém systému), děje se to, že při opětovném zadání vstupu A dostaneme výsledek A', B', C' etc., přičemž bude platit, že pokud je síť odladěná, měly by všechny výsledky splňovat zadání ke spokojenosti provozovatele i zákazníka. Opět zdůrazňujeme funkci produktové variace, která navíc slouží jako ochrana před porušením autorského práva.

2 Gellova agenční teorie a její aplikace na vizuálně-generativní AI

Z agenční teorie Alfreda Gella zavedeme několik pojmů. Jednak je to samotný pojem *agence* – aktivní působení na druhého s nějakým záměrem. (Gell, 1998, str. 16) Z něj plynou pojmy *agent* a *pacient*, tedy na jedné straně ten, kdo svou agencí aktuálně působí na druhého, na druhé ten, kdo toto působení aktuálně přijímá.[7] To, prostřednictvím čeho agent v uměleckém kontextu na pacienta působí, je zváno *index* neboli konkrétní (vizuální) artefakt. Základní vztah v rámci nexu sociální umělecké agence tedy vypadá tak, že index je nositelem agence, kterou do něj vtiskl agent k oslovení pacienta.[8] Pacient nahlíží, že je na něj umělecky působeno, a doptává se sám sebe na povahu agence, tedy na to, co asi kdo zamýšlel tím, že mu toto dal před oči. Gell toto tázání nazývá *abdukcí agence.* (Gell, 1998, str. 14) Jedná se o nepravou indukci založenou na domněnce nebo vágním pocitu z vizuálního artefaktu. Pokaždé, když se ptáme, kdo mohl konkrétní objekt vytvořit nebo pro koho nebo co byl zamýšlen, provádíme tuto abdukci nezávisle na tom, jestli existuje odpověď na tyto otázky nebo jsou-li tyto otázky k tomuto objektu relevantní.

Gellova teorie navíc umožňuje pojmout za agenty (a pacienty) i neživé objekty, např. antropomorfizované hračky, vozidla, budovy (Gell, 1998, str. 18) nebo tvořivé technologické artefakty typu VGAI. Agentem či pacientem se stává prostě to, co vstupuje do agenčního vztahu, tedy co přijímá agenci nebo čemu je agence přisuzována abdukcí.

Pokud představený systém přizpůsobíme naší výzkumné situaci, dojdeme k závěru, že v rámci Gellovy agenční teorie není problém přisoudit jakékoliv VGAI uměleckou agenci a jakémukoliv jejímu výtvoru ráz indexu, jehož vznik je navíc vlivem stochastiky vlastní VGAI samotné alespoň částečně nepředvídatelný. O síti samotné ovšem budeme hovořit jako o tzv. *sekundárním* agentovi.

Sekundární agence je umožněna vágní povahou abdukce: nezáleží na tom, jestli pacient atribuuje agenční záměr *řádně*, tzn. jestli objekt abdukce skutečně je intendujícím primárním agentem. Sekundární agent je takový agent, který esteticky působí na pacienta bez toho, aby cokoliv sám intendoval, tedy měl vlastní záměry podmíněné vědomím, vůlí apod. Gell (Gell, 1998, str. 21) používá příklad minometčíka a miny: ačkoliv je to mina, která aktuálně působí na pacienta, jedná se o sekundární agenci, jelikož mina nemůže intendovat. Intendující primární agent je minometčík, který minu pokládá se záměrem zabít. Přesto říkáme, že vojáka „zabila mina", nikoliv že „jej zabil minometčík", pracujíce zcela srozumitelně se sekundární agenční atribucí, aniž by kdo snad podezříval miny samotné z nekalých úmyslů.

Vycházíme zde z předpokladu, že neuronové sítě nemají vědomí a nemohou tedy intendovat. Jejich působení na pacienta skrze vygenerované obrázky (indexy)[9] má ráz plnění rozkazu nebo zadání. Za intenci, s jakou VGAI tvoří, tedy můžeme a musíme dosadit intence jejího autora, programátora, provozovatele a uživatele, tedy obecně primárního agenta: „[sekundární agenti] jsou objektivní ztělesnění moci nebo možnosti zamýšlet jejich užívání". (Gell, 1998, str. 21, autorský překlad) Sekundární agenti jsou takové předměty, jejichž užívání je spojeno se záměrem na straně uživatele. Pacient se sice abduktivně vztahuje k VGAI (se stejnou personifikační figurou jako v předešlém příkladu k mině, ptá se „co VGAI zamýšlela, když přede mne postavila tento obrázek"), ví-li, že hledí na vygenerovaný obrázek, ale v dalším kroku své tázání nutně musí usměrnit na tázání po primární agenci („jaký záměr měl původce sítě nebo ten, kdo ji využil k získání tohoto obrázku"). Platí tedy, že kdokoliv, kdo plánuje VGAI vytvořit nebo využít jakýmkoliv způsobem, vtiskává svou agenci do ní jako do nástroje a skrz ní do jejího produktu, cílového indexu. Zdálo by se tedy, že z intencionálního hlediska je VGAI *průchozí*, nepřidává nic ke vztahu agence a pacience, které volně proudí skrz její indexy. V následující kapitole se pokusíme ukázat, že užívání AI je ve všech svých instancích spojeno s intencí *ošálit* a že zmiňovaná intencionální průchodnost je pouze zdánlivá.

[7] Agence a pacience nejsou fixní kategorie, mimo aktuální agenční vztah věci a lidé zůstávají pouze potenciálními agenty nebo pacienty.

[8] Index je tedy v druhém plánu také pacientem, protože rovněž přijímá agenční působení agenta.

[9] VGAI má tu specifickou vlastnost, že se jedná o index a sekundárního agenta zároveň, o „tvořivou věc", jejíž výtvory jsou k síti samotné ve vztahu pacience.

3 Závěr – falzifikační povaha agence

VGAI je sice kuriózní svým jedinečným charakterem „tvořivé věci“, protože její produkty jsou unikátní vizuální díla. Jelikož však nemůže konkrétní podobu svých výstupů intendovat, jedná se o sekundárního agenta, který k sobě přibírá více primárních agentů (provozovatele, programátora, uživatele) a indexy, které generuje, tak syntetizují intence všech relevantních primárních agentů. Jejich vytrasování není snadné a v tuto chvíli ani podstatné.

Zeptáme se znovu, co obecně zamýšlí primární agent, který za svůj nástroj volí jako sekundárního agenta VGAI. Domníváme se, že jedna intence je vlastní užití této techniky jako takové, jako je užívání kladiva vlastní intence zatloukání, a nezáleží na tom, jestli hřebíků nebo suků. Sekundární agent je objektovým ztělesněním možnosti být užit se záměrem[10] a my se tážeme po obecné povaze tohoto záměru. Tímto univerzálním záměrem je podle nás *zastírající snaha přesvědčit pacienta, že má co do činění s lidským výtvorem* a že svou vizuální zkušenost může modelovat podle svých dalších zkušeností s konvenčními produkty lidské umělecko-řemeslné práce.

VGAI, jak je v době psaní příspěvku tržně implementována, je v první řadě služba, která na zadání generuje produkt s užitnou hodnotou. Tomu odpovídá způsob práce s tréninkovou databází a asimilace stylů: obchodním cílem je, aby VGAI generovala co možná nejuspokojivější reprezentace slovních zadání. Nejuspokojivější uměleckou reprezentaci slovního zadání tvoří lidský umělec. Produkty VGAI se tedy stylem nutně budou blížit pracím lidských umělců, na kterých je právě proto VGAI trénováno. Stylová návaznost na lidské vizuální umění je zcela záměrná a z obchodního hlediska nepostradatelná. VGAI generuje reprezentace nebo exempláře stylů, v nichž tvoří živí umělci, a odvolává se přitom na ustavené interpretační a abduktivní techniky lidské vizuální kultury. Řečeno jinými slovy, produkt VGAI se tváří jako lidský výtvor a žádá, aby s ním tak bylo při pohledu na něj nakládáno, ačkoliv jím není. Samotná nutnost připojování vysvětlujících popisků „použitý obrázek byl vygenerován umělou inteligencí“ poukazuje na relativní úspěšnost zastírající snahy, která z výše zmíněných důvodů musí být abduktivně atribuována autorům i uživatelům VGAI. Je to proto, že VGAI jako sekundární agent pouze facilituje agenci primárního agenta, jehož intence je v končeném důsledku otištěna v artefaktu. Průvodní snahou primárního agenta je, aby si pacient buď neuvědomil, že obrázek je vygenerován, nebo aby na to nebral zřetel a kognitivně s ním nakládal jako s lidským výtvorem, tedy aby jej dekódoval podle kódu lidského vizuálního umění. Řečeno pojmy teorie agence, primární agent pacienta skrz index záměrně *pobízí k nesprávné abdukci,* která by s indexem pracovala v irelevantních kategoriích „umělcova záměru“ a „reprezentace skutečnosti“.

Posledním krokem příspěvku vystoupíme z agenční teorie do teorie umění, odkud si přisvojíme jeden klíčový termín. Uměnověda totiž zná snahu přimět indexem pacienta k nesprávné abdukci agence a k nesprávnému dekódování založenou na zastírání, říká jí *falzifikace*. Nejde nám o falzifikaci v podobě zhotovení kopie a vydávání této kopie za originální dílo. Jak jsme zdůraznili v úvodu, VGAI nekopíruje, nýbrž variuje databázovaný obsah. Našemu případu lépe odpovídá kontext toho falza, které jeho autor pouze vyvede ve *stylu* napodobovaného autora a následně jej opatří falešným podpisem nebo značkou pro dodání věrohodnosti. Domníváme se, že schéma práce s VGAI v mnoha svých rysech odpovídá takto popsané falzifikační činnosti, ovšem uchopené trochu šířeji a abstraktněji. To, co se napodobuje, je styl, s jakým lidští umělci přistupují k reprezentaci reality. Výsledný produkt se s žádným existujícím dílem neshoduje, pouze některá připomíná natolik, aby se vedle nich mohl zařadit. Svou podobou apeluje na to, aby byl vnímán jako tato díla, nikoliv jako syntetická reprezentace stylu těchto děl. Přitom se nejedná o reprezentaci žádné prožité reality, pouze o nápodobu výsledku tvůrčího procesu bez toho, aniž by musel tento proces být proveden.

Naším výstupem není to, že každá falzifikace je špatná, proto je VGAI špatná. Jak dokazuje praxe vystavování falzifikátů *jako* falzifikátů vedle jejich originálů,[11] falzum není neumělecké. Co falzifikát činí potenciálně sdělným je však právě jeho odhalení a demaskování. Náležité zhodnocení přínosu VGAI a generovaných indexů na poli estetiky[12] nemůže a nebude opomíjet atribuci falzifikační agence primárním agentům, kteří s VGAI nakládají. V každé relační struktuře obsahující takto generované indexy bude pacient nucen nejprve přijmout, že měl být oklamán.

[10] Viz předchozí citace

[11] Příkladem je výstava *Falza? Falza!* Kurátorky Olgy Kotkové konaná ve Šternberském paláci od 19. 11. 2021 do 1. 5. 2022, viz [6].

[12] např. schopnosti VGAI nepřímo vytvářet skrz interakci s uměleckým světem hybridní styly nebo dokonce odhadovat budoucí trendy z předešlého vývoje

Literatura

Alfred Gell: *Art and Agency, an Anthropological Theory*, Clarendon Press, Oxford, 1998

J. M. Corchado, S. López F., J.M. Núñez V., R. Garcia S., Pablo Chamoso (2023): *Generative Artificial Intelligence: Fundamentals.* ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 12(1), e31704. https://doi.org/10.14201/adcaij.31704

Dave Bergman, Cole Stryker: *What are diffusion models?* [Online] Dostupné z: https://www.ibm.com/think/topics/diffusion-models

Q. Yao: *Application of Artificial Intelligence Virtual Image Technology in Photography Art Creation Under Deep Learning*, in *IEEE Access*, vol. 13, pp. 14542-14556, 2025, doi:10.1109/ACCESS.2025.3529521, dostupné z: https://ieeexplore.ieee.org/document/10840187

Jos de Mul: The work of art in the age of digital recombination. In *Digital Material: Tracing New Media in Everyday Life and Technology* (Marianne van den Boomen, Sybille Lammes, Ann-Sophie Lehmann, Joost Raessens, and Mirko Tobias Schäfer eds.), Amsterdam University Press, 2009, dostupné z: http://www.jstor.org/stable/j.ctt46mxjv.9

Olga Kotková (ed.): *Falza? Falza!,* Praha: Národní galerie v Praze, 2021, ISBN 978-80-7035-782-8.

Dave Bergman, Cole Stryker: *What are diffusion models?* [Online] Dostupné z: https://www.ibm.com/think/topics/diffusion-models

Mind Wandering during Lectures

Magda Stanová

Academy of Fine Arts in Prague
U Akademie 4, 170 00 Prague, Czech Republic
magdalena.stanova@avu.cz

Abstract

Lectures, panel discussions, and conferences are formats for collective listening, but they have developed conventions that make perception difficult. In this drawing-based talk, I will dissect the decorum of academic presentation formats and outline how the attention of the members of the audience diverges and converges with that of the lecturer. I will compare cognitive advantages and disadvantages of various presentation tools from the present and the past. I will also suggest that once a piece of research has been reduced to a text format, it might be difficult to add other modes to it later without them being auxiliary, and that, if we don't pick our presentation tools, the tools we use might be nudging us into doing research that fits them.

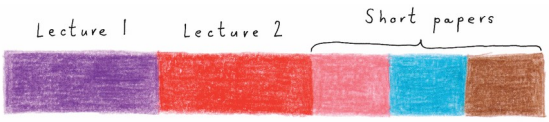
1 This is not an oral paper

In "Singing Euclid: the oral character of Greek geometry" [1], Viktor Blåsjö argues that Euclid's *Elements* wasn't written as a text but as a recording of spoken presentations —"Writing was merely the oral explanation put down on paper (or papyrus, rather)", and adds that the medium of text wasn't used yet to its full potential.

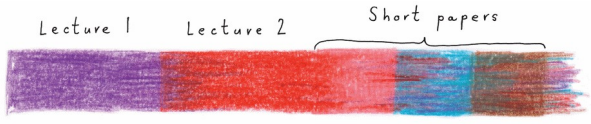
Today it's the other way around. Text became so dominant in academia that even conference presentations are called "papers" or "oral papers". The properties of text spill into talks. I've prepared my contribution to the conference as a talk, not as a paper, and I want to present it in these proceedings as such. Therefore, it is written as if it was spoken, and the images are embedded directly into sentences, since drawings and speech are like two legs that move the lecture forward.

2 Presentation conventions

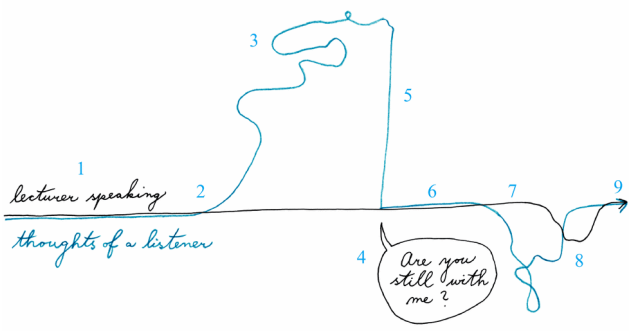
The program of conferences looks something like this



with a clear distinction between where one presentation ends and where another presentation starts. But when presentations are interesting, you keep thinking about them even when another speaker is already speaking. So, in our minds it looks more like this:



And something like this happens also during single lectures:



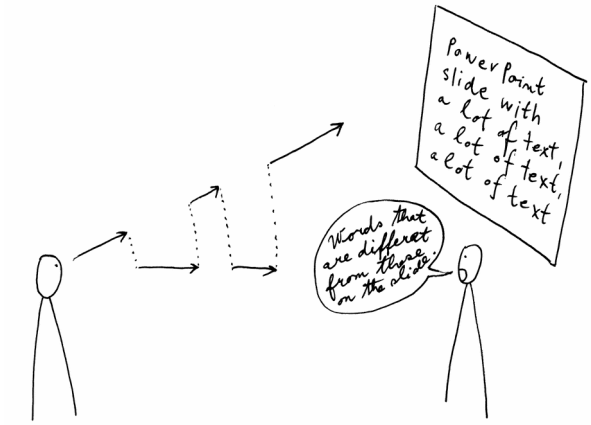
1 A lecturer is giving a lecture, so her mind is hopefully paying attention to what she is saying. A listener, at least at the beginning, is also paying attention, so they are both thinking about the same thing at the same time. But then 2 maybe the lecturer says something so interesting that it will 3 spark the imagination of the listener and their thoughts will diverge from those of the speaker. Until the speaker notices it and says 4 "Are you still with me?", which interrupts 5 the line of thoughts of the listener and brings them back to those of the speaker.

And then the lecture continues 6, and maybe after a while the lecture gets too boring 7 or too complex, so the listener gets lost again. And occasionally, especially when it looks like nobody is paying attention anymore, the speaker gets disconcerted and loses her attention as well 8, which is awkward and, unfortunately for the speaker, usually attracts the attention of the listeners. And, for situations like this, the artist Daniel Yovino recommends [4] to have ready a card with an emergency conclusion, which will help 9 to bring the lecture to the end.

Conferences are formats for collective listening, but they took on conventions that make listening difficult: You sit quietly for a long time without moving, looking in one direction. The program is usually very dense, so there is little time to think about what has been said. Lecturers, especially in humanities, often read aloud complicated texts that are good for reading, but not for listening. Sometimes there is no good place to put notes or a text, so the lecturer ends up holding them in hand. In some settings, you must hold a microphone in the other hand, so you end up having no hand left for gesturing.

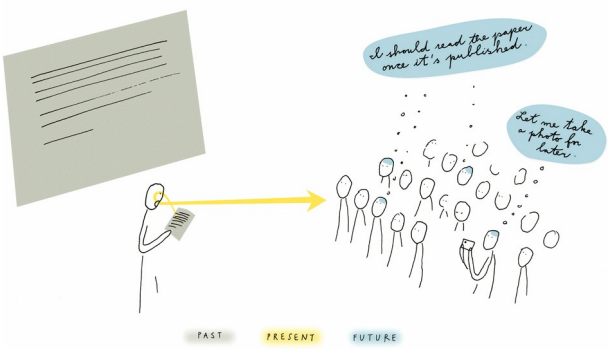


When you're presenting, it feels like you have to have a PowerPoint presentation. Slides sometimes include texts that are different from what the speaker is saying, so, as a listener, you don't know if you should keep listening to the speaker or read the text on a slide.

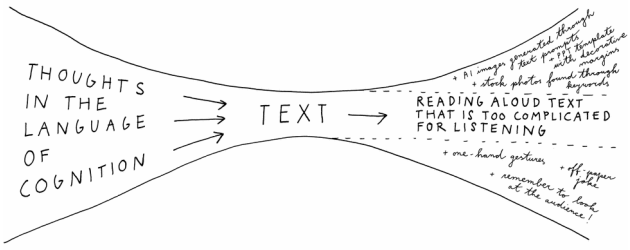


Also, lecturers tend to click through slides very quickly, as if the eyes of the listeners were photo cameras able to perceive everything within one blink.

In academia, there is a preference for text. When an image gets shown, it is rarely trusted by itself, and a verbal explanation is added. Journal articles are valued more than presentations at conferences. Lectures are live readings of future articles. Instead of being together in the present moment, the lecturer reads a text prepared in the past, while the listeners are deciding whether they will read the text when it's published.



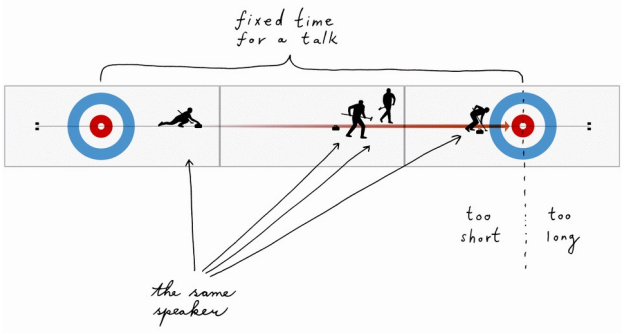
Unlike text, the lecture is a multimodal format, so besides the verbal mode, there are also intonation, gestures, images, sounds, facial expressions. But when a piece of research has been at first reduced to the verbal mode, it might be difficult to add other modes later without them being auxiliary. To convert text into a lecture, we could subsequently add:



+ images for illustrative purposes like stock photos found through keywords or AI images generated through text prompts
+ a PowerPoint template with decorative margins
+ one-hand gestures (if you have a hand left)
+ maybe an off-paper joke
+ and don't forget to look from time to time at the audience!

The oral paper has become multimodal, but it's an awkward multimodality.

Many conferences are strict about timing and this aspect of giving a lecture reminds me of curling.



They give you, let's say, 20 minutes. Now, you don't want to finish too early, because then you look like you have not much to say. But, if you pass the time limit, then they may cut you off and you won't finish your argument.

For many lecturers the given amount of time is not enough, so they search for ways to squeeze as much material as possible. One of the techniques is to write the whole lecture down, and, as the end of the lecture is approaching, read the text as fast as possible. The lecturer is satisfied that he said everything he wanted to say, but the listeners weren't able to perceive it.

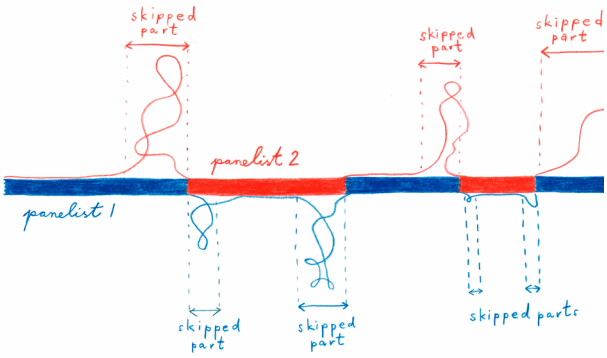
Now let's look into panel discussions. Here we have two or more people speaking one after another.



Ideally, the panelists would have time to prepare their reactions to what other panelists have said.



But silence during panel discussions is considered awkward—in art history terms, we could say that there is *horror vacui* of silence. So, panelists talk right after each other, which means that they have to prepare their reactions while the other panelists are still speaking.



This can lead to embarrassing situations when a panelist later asks about something that has been already answered.

Not paying attention is usually considered a bad thing. But there are various reasons for not paying attention. It's like when you read a book. Sometimes you zone out because you're too tired or because something important has happened during your day, so you keep thinking about it. But sometimes you stop reading because you start to ponder about what you've just read. Reading doesn't consist only of reading itself but also of thinking about what you've read.

When you are not paying attention during a lecture because you have started to think about something that you've heard or seen during the lecture, it is a good thing.

3 Time for mind wandering, fixed or flexible, in or out of touch

Now let's compare various presentation formats and tools to see if and how they provide for mind wandering, whether they allow improvisation, and if they reinforce or surpress the pointing gesture.

Outside academia, there are genres where time for mind wandering is already embedded in the format. Radio interviews (but not podcasts) are intermitted by songs after about every 10–15 minutes of speaking. Within songs, there is a refrain—a part that repeats itself, so you don't have to pay attention to it as much. Movies end with closing credits which gives you time to sit in the dark with music for several minutes and let the movie sink in. And then the Catholic church has a genre called the way of cross, where there are several stations scattered around a hill, usually little chapels, and in each station, there is one "slide" (if we would call it in conference jargon), usually in the form of a painting or a relief. Participants walk from one "slide" to the next "slide", so the time for mind wandering is connected with physical wandering.

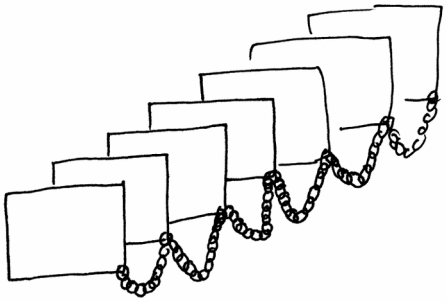
But even within academia, before the spread of PowerPoint, there were presentation tools that provided time for mind wandering. When you used a chalkboard (also called blackboard), you had to clean it, which took quite a lot of time. And, if you did it quickly, there was another advantage to it: the board held a memory of what had been there before. The most recent notes were sharp, the older ones were fuzzy, the transition between them was slow, much slower than the transition between the slides in PowerPoint.

Chalkboards were replaced by whiteboards, as if they were the same thing, as if the whiteboard was an equivalent of the chalkboard. But there are significant differences when it comes to their use properties, the aesthetics, and the dynamics of the presentation. The chalk allows you to draw lines of various thickness: you can write with the end or with the side, and you can adjust the thickness by breaking the chalk. The chalk also responds to how strong you push (like the piano), so you can make a gradient. With a marker, you can only draw with its tip.

The surface of the chalkboard is a bit rough, so when you write, it invites you to pause, you can lean on it with the chalk, so your arm can rest while you decide how to continue. This also creates a bit of suspense when your audience tries to predict what's going to come next. This is unlike writing on the whiteboard, the surface of which is slippery, so it provides no time to rest.

Writing with the chalk was slow, so you wrote only what was necessary. The chalk forced you to be concise. The computer keyboard makes typing too easy, which makes the PowerPoint slides wordy. On the contrary, drawing with chalk was easy. Drawing with the keyboard is impossible, and drawing with a mouse or a touchpad is slow and clumsy.

Before the spread of PowerPoint, lecturers used the overhead projector. Unlike in PowerPoint, where slides are bound together and, if during the lecture you decide to skip something, you have to quickly show everything that you wanted to skip,

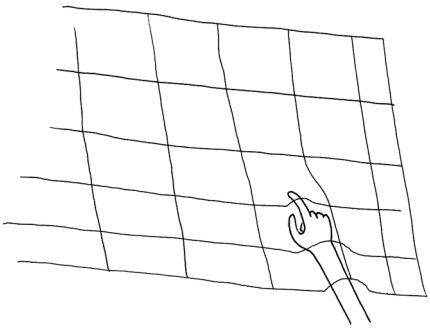


with the overhead projector you can improvise. You can add or skip slides without the audience knowing, you can draw live, and you can superpose multiple layers of transparencies, like Roger Penrose does [2] when he explains the Schrödinger's Cat thought experiment.

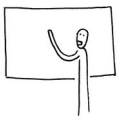
Also, with the overhead projector you can touch your slide, you can point your finger on it.



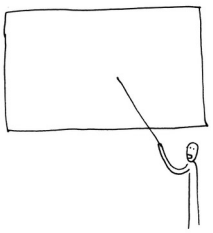
With the digital projector, you're not touching the slide. The slide is like a blanket that covers you.



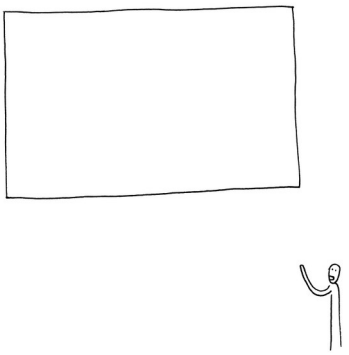
But even like that you can point only in small settings



because the bigger the speaker,



the bigger the lecture room,

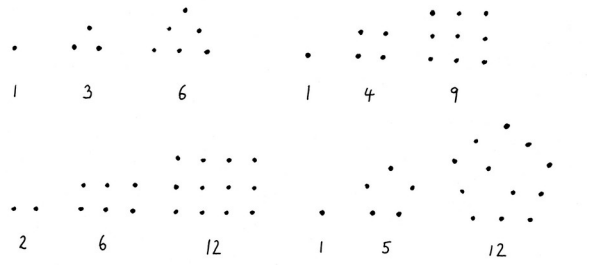


and your slides are getting out of touch.

4 Following tools

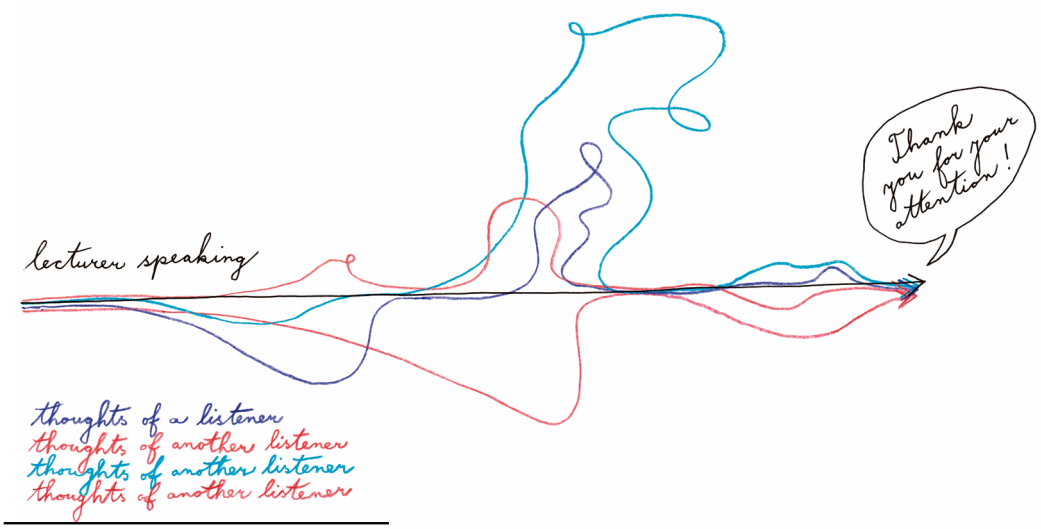
Presentation formats and tools change over time. Today we use PowerPoint and the whiteboard, 30 years ago we would use the overhead projector or the chalkboard, 200 years ago the magic lantern, and 2500 years ago we would

draw in sand, like Pythagoreans probably did.[1] Today researchers don't draw in sand, because it would feel too eccentric, and they (with the exception of Roger Penrose) don't use the overhead projector, because it would feel too retro. We tend to stick to what's currently the standard. But if Pythagoreans used PowerPoint, or even if they drew on a vertical surface like a chalkboard, they wouldn't be able to add little stones, and so they possibly wouldn't research figurate numbers.



Small stones catalyze ideas about figurate numbers better than other tools. If we don't pick our presentation tools, the tools we use might be nudging us into doing research that fits them.

So, thank you, in case you managed to pay attention, but also if you were mind-wandering.



[1] There is a nice drawing by Michal Škrovina in Milan Hejný's book *Geometria naučila človeka myslieť* (SPN, 1979) showing Pythagoreans drawing in sand. Raviel Netz in *The Shaping of Deduction in Greek Mathematics* (Cambridge University Press, 1999) argues that preparing sand for drawing would be too demanding, but the existence of ancient Greek vases from circa 450 BC where teachers and students draw with long sticks on the ground suggests otherwise (e.g. Louvre, collection number Cp 977, catalogue number G 448, tagged "scène d'école". https://collections.louvre.fr/en/ark:/53355/cl010270297)

References

Blåsjö, V. (2020). Singing Euclid: the oral character of Greek geometry. Podcast episode and blog post. In: Opiniated History of Mathematics (podcast) and Intellectual Mathematics (blog). Posted June 21, 2020, Available online: https://intellectualmathematics.com/blog/singing-euclid-the-oral-character-of-greek-geometry/.

Penrose, R. (2014). Consciousness and the Laws of Physics. Video. Filmed April 13, 2014. Posted October 2015 as "Sir Roger Penrose - Keynote Speech at Towards a Science of Consciousness 2014". Available online: https://www.youtube.com/watch?v=dFs-N-XFkQ0&.

Stanová, M. (2025). Mind Wandering During Lectures. Video paper. In: *Proceedings of the 1st symposium Forum Artistic Research: listen for beginnings*. https://doi.org/10.22501/rc.3847167

Yovino, D. (2008). How to Live Productively Using Your Hipster PDA. Video. Filmed November 29, 2008, at "The Art of How-To" in the Koret Education Center of SFMOMA, San Francisco. Posted May 20, 2010, by Daniel Plus Yovino on YouTube. https://www.youtube.com/watch?v=-HqMV7XNfS4.

Konceptuální redukce a smrt obrazu

Aleš Svoboda

Univerzita Karlova, Fakulta humanitních studií
Pátkova 2137/5, 182 00 Praha 8 – Libeň
ales.svoboda@fhs.cuni.cz

Abstrakt

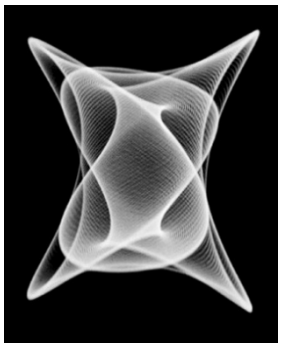
Zavedení strojů nového druhu, počítačů, od počátku vybízelo k uměleckému využití. Kreativní programátoři a experimentující umělci začali spontánně hledat způsoby, jak uvolnit předpokládané mentální schopnosti chladných kovových skříní. Od této doby nás dělí víc než 60 let. Počáteční rezervovanost k „počítačovému“ umění, přezíravost nebo dokonce vážné pochyby veřejného mínění o něm se aktuálně změnily spíše v obavy ze ztráty kontroly nad oblastí, která byla definičně lidská, nebo dokonce ve frustrace z vlastní inferiority. Je na místě přiznat strojům jejich vývojovou nadřazenost, nebo se zatím jedná jen o naši vyplašenou fantazii?

1 Zrod umění s využitím počítačů

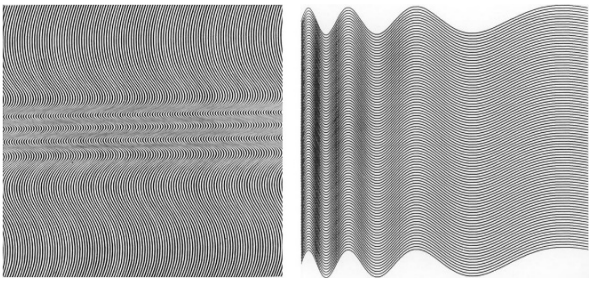
Objev strojů nového druhu, počítačů, respektive samotné úvahy o jejich zkonstruování, probudily záhy představy o schopnosti strojů myslet (Turing, 1950) a projevovat funkce živých organismů, jako je rozmnožování (Neumann, 1966) či vnímání a reagování (Wiener, 1948).

Praxe s počítači od počátku zaujala i jako předpokládaný pomocník umělecké kreativity. Dnes můžeme s poměrně dobrou znalostí dostatečně zmapovaného historického dění identifikovat na konci 50. a v první polovině 60. let tři zásadní počítačově-umělecké praxe (Dietrich, 1986): zájem o estetické působení periodických křivek (Obr. 1), napodobování stávajícího, aktuálně oceňovaného umění (Obr. 2) a modelování tvůrčího procesu, pochopeného jako proces logické volby mezi alternativami (Obr. 3).

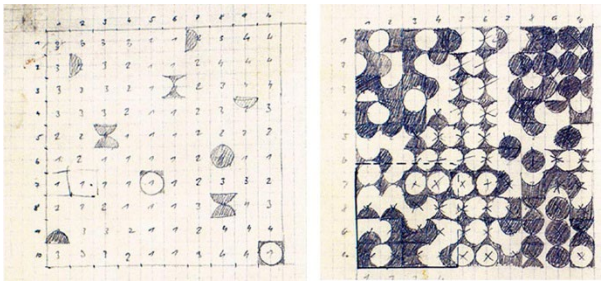
Samozřejmě v paralele k lineárnímu fungování počítačů se první pokusy jejich uměleckého využití týkaly literatury a hudby. Koincidence s poznáním síly statistických nástrojů vedla k zájmu o vyhodnocování pravděpodobnosti znaků v lineárních řetězcích (Shannon, 1964 a Moles, 1966)[1]. Dvojrozměrnost, případně i trojrozměrnost vizuality ovšem vyžadovala nástroje, které dokázaly porovnávat proměnné hodnoty napříč strukturami, což se zprvu zdálo jako nesnadný úkol. Adaptace „Hry života“ Johna Conwaye do vizuálních struktur v díle Paula Browna byla velmi účinnou, a nakonec i esteticky ceněnou cestou (Brown, 2008).



Obr. 1: Ben F. Laoský: Oscilón, 1953.



Obr. 2: Bridget Rileyová: Pád, 1963 a Michael Noll: Devadesát paralelních sinusoid s lineárně se zvětšující frekvencí (Variace na tvorbu Bridget Riley), (1964).

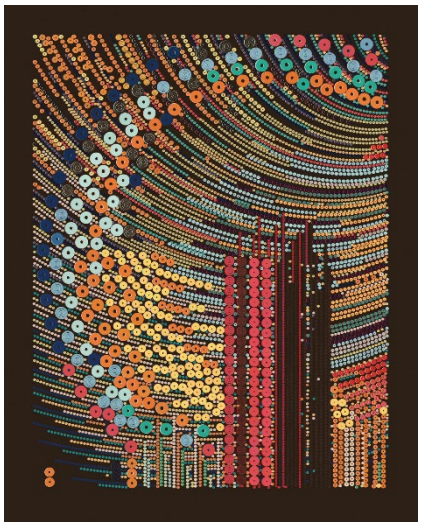


Obr. 3: Zdeněk Sýkora: Zadání a vizualizace programu pro černo-bílou strukturu, 1964

[1] V algoritmech působila vložená náhodnost jako projev umělecké intuice, usměrnění nesourodé libovolnosti pak bylo svěřováno pravděpodobnostním omezením.

2 Počítačové generativní výtvarné umění

Následný vývoj počítačového výtvarného umění jednoznačně směřoval k tomu, co se intuitivně nazývalo „generativní“ umění a přesné terminologické označení jako „počítačové generativní umění“ pak získalo především díky práci Philipa Galantera (Galanter, 2016).
Jeden z průkopníků počítačového umění, Herbert W. Franke, věnoval pozornost generativnímu umění v retrospektivním pohledu v roce 1989 (Franke). Tehdy ovšem zohledňoval propojení matematiky s uměním jako účinným nástrojem uplatňování řádu v uměleckých artefaktech. Druh takové procedury navrhuje nazývat „generativní matematikou“, která by se definovala jako „*…studium matematických operací vhodných pro generování uměleckých obrazů“*. Nicméně algoritmický základ umění vytvářeného počítačem se v druhé polovině 60. let překvapujícím způsobem shodoval s obecným přesměrováním uměleckého zájmu z tvorby intuitivních jedinečných artefaktů (vysvětlovaný tzv. metodou „ruka-oko“) na racionální formulování samotného postupu jejich tvorby v konceptuálním umění. Jinak řečeno, pokud konceptuální umění požaduje mentální, tedy rozumějme verbální formulaci uměleckého díla v přirozeném jazyce ještě před jeho fyzickou tvorbou[2], sestavení algoritmu díla v programovacím jazyce je tomu zcela adekvátní. V tomto smyslu také vymezuje Galanter „počítačové generativní umění“ jako podoblast širokého pole všeobecného generativního umění.



Obr. 4: Výstup z projektu QQL, generativního uměleckého projektu s neomezenou formou, vytvořený uživatelem hex6c v roce 2022.

Současná praxe počítačového generativního umění (Obr. 4) staví na postupech NFT (Non Fungible Token) a ukládání konstruujících algoritmů na úložištích „blockchainu“ jednak pro zajištění nepřenosné jedinečnosti uměleckého artefaktu, jednak pro možnost produkovat „cykly“ souborů typově podobných děl, jejichž okamžité vzezření je variabilní, překvapivé a individuálně vyžádané podle požadavku vlastníka díla (Franceschet, 2025).
Tak jako v průkopnických dobách počítačového umění i současné generativní umění sice rozvíjí tradiční uměleckou tvořivost, ale ta je stále standardně „dozorovaná“ umělcovou intencí. Zpětnovazební cyklus mezi zpočátku nejasným cílem a postupně objevovanými (případně zamítanými) dílčími objevy je řízen člověkem a nelze ho chápat jako projev umělé inteligence, byť kombinatorickou mohutnost tvorby dostává na novou úroveň, které by samotný umělec těžko dosáhl.

3 Umělá inteligence ve výtvarném umění, symbolická AI a LLM

Za velmi raný regulérní pokus rozvinout umělou inteligenci v umění se pokládá projekt *Aaron* Harolda Cohena (Cohen, 2002). Jedná se o typický příklad opírající se o symbolickou AI, dnes nazývanou GOFAI, tedy „starou dobrou umělou inteligenci“, kdy se umělec pokusil popsat celý repertoár používaných vizuálních prvků s jejich proměnnými charakteristikami a možnými vzájemnými vazbami, které při kombinatorickém generování vyplňují hypotetický Cohenův stylový prostor novými a novými variantami kompozic [3](Obr. 5).



Obr. 5: Harold Cohen: Počítačem generovaná kresba (ručně kolorovaná), 1974

K počítačovému postižení myšlení se zdály být užitečné dvě jasné intuice: jednak vytvoření dobře popsaného pole symbolických reprezentací spojených logickými pravidly, jednak napodobení činnosti centrální nervové soustavy. První je přístupná naší racionální představě fungování mysli, ovšem posléze se jevila jako nekonečná, respektive nedokončitelná idea. Při realizaci

[2] „*Myšlenka se stane strojem, který dělá umění.“* Sol LeWitt, „Odstavce o konceptuálním umění“, *Artforum* (Summer 1967, roč. 5, č. 10): 79-83.

[3] Harold Cohen úspěšně navázal na svou „nezobrazivou“ etapu i tvorbou figurativní, kterou aforisticky označoval jako zobrazivou malbu bez zrakového aparátu.

druhé intuice se procesy vědomí noří do tmy černé skříňky.
Ve 40. letech 20. století vytvořili W. S. McCulloch a jeho student W. Pitts model neuronu, v roce 1958 Frank Rosenblat sestavil první funkční síť, tzv. perceptron. Po kritice Marvina Minskeho a Seymoura Paperta se ani tato cesta výzkumu nezdála dostatečně perspektivní, nicméně v polovině 80. let k myšlence umělých neuronových sítí badatelé opět vrací; D. Rumelhart, G. E. Hinton a R. J. Williams ve studii *„Learning Internal Representation by Error Propagation“* (Rumelhart, 1986) přicházejí s novým algoritmem zpětného šíření chyby (tzv. backpropagation) pro trénování vícevrstvých neuronových sítí a pokládají tak základy rozvoji hlubokého učení.

4 Umělé neuronové sítě a GAN

Do povědomí široké veřejnosti se umělá inteligence dostává po roce 2015, po uvedení programu *DeepDream* od Googlu, který využil konvoluční neuronové sítě, tedy typ algoritmu hlubokého učení, k detekci a vylepšení tváří a vzorů v obrazech. Tato technologie se ukázala jako zvlášť účinná při analýze vizuálních obrazů, takže i umělci se začínají zajímat o možnosti jejich generování, posléze především pomocí generativních soupeřících sítí (GAN).
Projekt „Next Rembrandt/Příští Rembrandt“ je často citovaný příklad umělecké tvorby právě s použitím generativní soupeřící sítě. Tato technologie byla vyvinuta v roce 2014 Ianem Goodfellowem a jeho kolegy na Montrealské univerzitě. Můžeme ji v jistém smyslu přirovnat k myšlenkám střídavé produkce a kritického sebehodnocení prostřednictvím odpovídajících programů, která přišla vizionářům počítačového umění na mysl už v 80. letech minulého století. Principem GAN je soupeření dvou umělých neuronových sítí coby hry s nulovým součtem. Jedna síť je generátorem, který tvoří nové obrazy, druhá síť je diskriminátorem, a hodnotí stupeň jejich příbuznosti k původnímu souboru.
Projekt „Příští Rembrandt“ v roce 2015 iniciovala nizozemská banka ING, reklamní agentura J. Walter Thompson. S týmem počítačových vědců z Technické univerzity v Delftu spolupracovala muzea, která spravují díla Rembradta van Rijna – Mauritshuis v Den Haagu a Museum Het Rembrandthuis v Amsterodamu, a byl vytvořen datový soubor ze 346 známých Rembrandtových obrazů. Algoritmus stanovil jako nejpravděpodobnější Rembrandtův námět poloprofil bílého muže kolem 35 let, s plnovousem a kloboukem, v černých šatech s bílým límcem. Odpovídající obrazy byly analyzovány v jednotlivých fragmentech (tj. jednotlivé obličejové části) a algoritmy hlubokého učení a techniky rozpoznávání obličejů následně dokázaly stanovit „vzorce“, nejpravděpodobnější struktury odpovídající Rembrandtovu kompozičnímu postupu, technice štětcových tahů, výběru barev. GAN byla vycvičena i v texturních vlastnostech obrazů, aby 3D tiskárna mohla připravit také imitaci povrchu malby. Po 18 měsících byl představen výsledek – dosud neznámý „Rembrandtův“ obraz (Obr. 6).



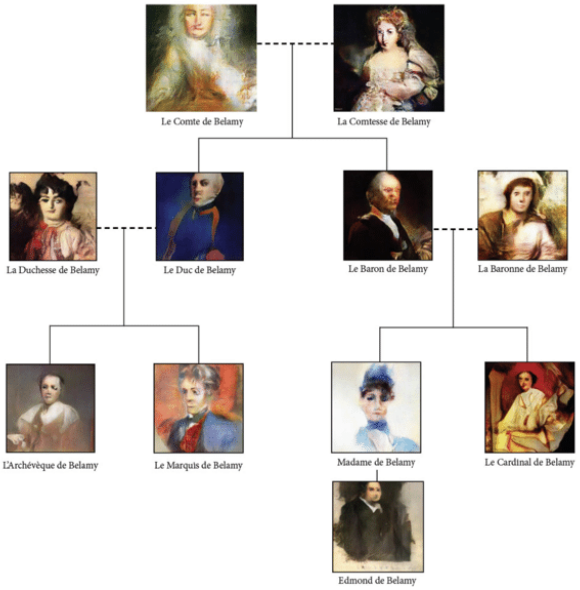
Obr. 6: The Next Rembrandt/Příští Rembrandt, 3D tištěná malba vytvořená pomocí algoritmů neuronové sítě, 2016.

Souhrnně lze tedy srovnat práci celého týmu a účasti GAN s prací padělatele. Výsledek je pouhým „kombinatorickým“ rozšířením stávajícího výtvarného pole a další přiřaditelné položky. Ostatně vzniklá technologie se dnes v tomto smyslu používá pro opravu poškozených děl, nebo doplnění jejich ztracených částí.
V roce 2018 vycvičila generativní soupeřící sítě francouzská skupina *Obvious* na 15 000 klasických malířských portrétech ze 14. až 20. století. Algoritmus pak dokázal produkovat soubor různorodých variací na typy portrétů. Z výsledků sestavil kolektiv celou fiktivní rodinu, seřazenou podle atributů módy. Edmond de Belamy se posouvá časově blíže ke 20. století.



Obr. 7: Obvious, Portrét Edmonda de Belamy, 2018.

La Famille de Belamy, rodokmen imaginárního šlechtického rodu, představuje souboru portrétů, které diskriminátor uznal jako relativně příbuzné konvencím jistých portrétních stylů (Obr. 8). Každý portrét obsahuje jistou míru zkreslení a pro strojové pravděpodobnostní postupy logickou míru neurčitosti realistického pojetí. Popravdě řečeno, v projektu vítězí přímočará umělecká spektakulárnost nad další možnou vrstvou genetické podobnosti fyziognomie portrétů, která by mohla být také součástí učení sítě.



Obr. 8: Rodokmen rodiny de Belamy, 2018.

5 Generativní umělá inteligence a promptové pokyny

Lev Manovich v elektronické knize *Artificial Aesthetics: Generative AI, Art and Visual Media*, kterou sestavil společně s Emanuelem Ariellim (Manovich – Arielli, 2024), hned v první kapitole uvádí exemplární případ užití generativní soupeřící sítě. Pro vytvoření tradiční čínské krajinomalby ji využila v roce 2020 studentka Princetonské university Alice Xue, která síť trénovala rozsáhlým datovým vstupem čínských štětcových krajinomaleb (téměř 2200 maleb ze 4 univerzitních a muzejních sbírek). Svůj postup nazvala model Sketch-And-Paint GAN (SAPGAN), a skládal se ze dvou GAN: SketchGAN pro generování map hran a PaintGAN pro následný převod z map hran na malbu. Došla k závěru, že v jejím případě se nejedná o podmíněný vstup, který generuje neoriginální výstupy, ale o režim bez podmíněného vstupu, který naopak generuje výstupy jedinečné, což má potvrzovat jí provedená varianta Turingova testu, kdy se 55 % z 252 respondentů přiklonilo k hodnocení generovaných děl jako výsledků lidské práce (Xue, 2020).



Obr. 9: Simulace čínské krajinomalby, vytvořená pomocí GAN v roce 2020 Alicí Xue.

Nehledě na nepříliš významný výsledek testu zdá se mi poněkud zavádějící pokládat snahu respondenta nezadat si při rozlišování příslušnosti hodnoceného díla ke stylu dané čínské krajinomalby jako důkaz originality. Podobně jako *Next Rembrandt* je cílem i očekávaným výsledkem pravděpodobností postižení přípustných formálních variant, ostatně jak tomu bývalo podobně bez užití umělých neuronových sítí v textových analýzách 60. a 70. let. Nová díla jsou skutečně nová, ještě nikdy předtím nevytvořená, ale jejich „typ“ je dán již dříve dobře definovaným postupem. V klasifikaci Margaret Bodenové se jedná o „kombinační“ kreativitu (Boden, 2016).

Manovich ve výše zmíněné knize uvádí dva příznačné experimenty s obrazovými AI generátory. Úspěch velkých jazykových modelů vedl k propojení slovních popisů s postupy rozpoznávání vizuálních struktur a tím k omračující možnosti vyvolat dostatečně adekvátní obrazový výsledek slovní formulací, „promptem“. Popis se může týkat nejen obrazových forem, kvalit a proměnných, ale i dostatečně zavedených stylových charakteristik ať už individuálních (Rembrandt, Bosch, Malevič apod.) nebo slohů, stylů skupin, uměleckých směrů (holandská malba 17. století, pop art apod.). Manovich zadal programu Dall-E-3 zadání „Sada počítačových komponent ve stylu holandského zátiší ze 17. století“. Kombinační přizpůsobení je velmi přesvědčivé (Obr. 10).



Obr. 10: Lev Manovich, PC-set ve stylu holandského zátiší ze XVII. století.

Podobně pokud máme určitou vizuální zkušenost se souborem Malevičových děl a zároveň s minuciózním stylem Hieronyma Bosche, shledáváme obrázek vygenerovaný v Midjourney na základě zadání „obraz od Maleviče a Bosche“ jako plnohodnotnou extrapolaci (Obr. 11)[4].



Obr. 11: Lev Manovich, Obraz od Maleviče a Bosche, 2022.

Oba příklady dokládají nepřímo dosavadní omezení AI generátorů – formulace promptů se musí opírat o jasné konvence zavedených vizuálních jevů. Čím obecnější a nekonkrétnější prompt, tím prázdnější a bezobsažnější je vizuální výsledek.

[4] Schopnost AI extrapolovat nejen v průsečíku dvou individuálních stylů (Malevič, Bosch), ale dokonce v rámci různých tvůrčích životních etap konkrétního malíře (suprematistická a geometricky stylizovaná realistická) je opravdu udivující.

6 Závěr

Stále se potvrzuje, že praxe obrazových AI generátorů, jako je např. Dall-E, Midjourney, DeeVid, Leonardo, Flux bohužel pouze projevuje schopnost kombinační kreativity, nikoliv kreativity transformační (Boden, 2016).

Ovládání prompty odkazuje k věcnému, obsahovému, kontextovému či stylovému popisu obrazu. Jinými slovy, jazykové postižení obrazu je již ze své podstaty invalidní pro hluboké a celistvé postižení účinnosti obrazu. Vizuální sdělení není redukovatelné na verbální popis, ostatně autonomie tohoto informačního kanálu je trvale potvrzována jeho udržováním, zušlechťováním a používáním. Při redukci obrazu na jeho slovní popis nastává jeho „umrtvení“, tedy emfaticky nazváno: smrt obrazu. Viditelné je formou „názoru“, ze kterého jakýkoliv prompt selektuje jenom dílčí postřehy, jež nezvládají popsat, nebo dokonce falzifikují „estetickou“ účinnost obrazu v původním smyslu Baumgartenovy definice, tedy jeho bezprostřední intuitivní vyvolání celé komplexní reakce vizuálního aparátu.

Je hypoteticky možné uvažovat o budoucím enormním zvětšení rozsahu modelů generativních neuronových sítí při jejich stejně obrovském navýšení a trénování na „velkých datech“ odpovídajících totálnímu shrnutí celého kulturního dědictví. Obávám se ale, že ani materiální, technické podmínky (kapacitní a energetické), a ani metodické možnosti (co všechno a jak) ke zvládnutí takového úkolu nejsou v dohledu. Algoritmizace nesourodých slepenců může vést opět jen ke kompilátům, bez výhledu na transformační kreativitu. A poslední, avšak velmi důležitá poznámka: veškeré výsledky nové, revoluční kreativity byly vždy zpočátku nesrozumitelné (v Turingově testu by neobstály), byly zdrojem pro následnou intelektuální práci, tříbení a hledání významu. Konec konců skutečný posun v tvorbě významu je záležitostí „lidského učení“, nikoliv „učení strojů“.

Poděkování

Tento výstup vznikl v rámci programu Cooperatio, vědní oblasti Vědy o umění a kultuře.

Literatura

Boden, M. A. (2016). *AI, Its nature and future*, Oxford: Oxford university press, str. 68–69

Brawn, P. (2008). From System Art to Artificial Life: Early Generative Art at the Slade School of Fine Art. In BROWN P. a kol. (eds.), *White Heat Cold Logic,*

British Computer Art 1960-1980, Cambridge, Massachusette – London, England: The MIT Press, str. 275–289

Cohen, H. (2002). A Self-Defining Game for One Player: On the Nature of Creativity and the Possibility of Creative Computer Programs, *Leonardo*, 35(1), str. 59–64

Dietrich, F. (1986). Visual intelligence: The First Decade of Computer Art (1965–1975), *Leonardo*, 19(2), str. 159–169

Franceschet, M. (2025). Modern Forms of Generative Art, *Leonardo*, 58(4), str. 397–402

Franke, H. W. (1989). Mathematics as an artistic-generative principle, *Leonardo*, 2(Supplemental Issue), Computer Art in Context: SIGGRAPH '89 Art Show Catalog, str. 25–26

Galanter, P. (2016). Generative Art Theory *in* Paul, Ch. (ed.), *A Companion to Digital Art*, Chichester: John Wiley & Sons, Inc.

Manovich, L. a Arielli, E. (2025). *Generative AI, Art and Visual Media*, https://manovich.net/content/04-projects/186-artificial-aesthetics/manovich_and_arielli.artificial_aesthetics.all_chapters_final.pdf [cit. 2026-04-27]

Moles, A. (1966). Information Theory and Esthetic Perception, Urbana a Londýn, University of Illinois Press

Neumann, J. von a Burks, A. W. (1966). *Theory of Self-Reproducing Automata*. Urbana a Londýn, University of Illinois Press

Rumelhart, D. E., Hinton, G. E. a Williams, R. J. (1986). Learning Internal Representations by Error Propagation, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Vol. I, D. E. Rumelhart and J. L. McClelland (Eds.) Cambridge, MA: MIT Press, str. 318–362

Shannon, C. E. a Weaver, W. (1964). *The Mathematical Theory of Communication*, Urbana, The University of Illinois Press

Turing, A. M., (1950). Computing Machinery and Intelligence, *Mind*, 59(236), str. 433–460

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine.* Paříž, Hermann & Cie, a Camb. Mass., MIT Press

Xue, A. (2020). End-to-End Chinese Landscape Painting Creation Using Generative Adversarial Networks, arXiv:2011.05552v1 [cs.CV] 11 Nov 2020

Lidé vs. Algoritmy: Preferujeme stále ještě lidské umění?

Julie Šišková, Noemi Chrásková, Laura Eklová, Filip Děchtěrenko

FF UK, Náměstí Jana Palacha 2, Praha 1, 116 38
Email: julie.siskova11@gmail.com

Abstrakt

S rozvojem umění generovaného umělou inteligencí můžeme ve společnosti pozorovat narůstající obavy z nahrazení umění produkovaného člověkem. Cílem naší studie bylo ověřit, zda přetrvává preference umění vytvořeného lidmi, konkrétně poezie. Ve výzkumu byly použity dvě sady básní, jedna sada vytvořená lidskými umělci, druhá sada vytvořená generativní umělou inteligencí (ChatGPT od firmy OpenAI). V experimentu jsme manipulovali s proměnnou autorství básní. Stanovili jsme dvě hypotézy. První předpokládala, že participanti budou básním s domnělým AI autorstvím přisuzovat nižší uměleckou hodnotu. Druhá hypotéza dále předpokládala, že domnělé autorství bude mít na hodnocení básně větší vliv než autorství skutečné. Zvolili jsme 2×2 between-subjects design: básně byly buď lidského, nebo strojového původu, a byly prezentovány jako vytvořené člověkem, nebo umělou inteligencí. Tím vznikly čtyři podmínky, do nichž byli participanti náhodně zařazeni. Každá báseň byla hodnocena pomocí Likertovy škály v šesti kategoriích. Výsledky statistické analýzy ukazují, že u básní vytvořených umělou inteligencí se objevil signifikantní rozdíl v hodnocení v závislosti na domnělém autorství. Nejhůře byly hodnoceny básně, u nichž bylo jak skutečné, tak domnělé autorství strojové. U básní vytvořených lidmi se tento efekt nepotvrdil, tedy skutečně lidské básně prezentované jako vygenerované AI nebyly hodnoceny statisticky hůře než básně prezentované jako lidské. Z výsledků lze tedy usuzovat na preferenci člověkem vytvořené poezie před básněmi vygenerovanými pomocí AI.

1 Úvod

Rozvoj umělé inteligence s sebou kromě rozvoje primárně analytických, výpočetních nebo mechanických AI přináší rovněž rozmach počítačem generovaného umění. V současné době dokáže AI napodobovat umělecké styly známých umělců (Aras, 2025), komponovat hudbu (Beatoven.ai, n.d.) nebo psát poezii (Porter a Machery, 2024). Oproti oněm více analyticky zaměřeným schopnostem AI však generované umění vyvolává i negativní ohlasy ve společnosti. Od protestů vycházejících z uměleckých kruhů až po více existenciální stesky, které lze najít např. v podobě komentářů pod YouTube videem AI vygenerované písně Daddy's car (Vincent, 2016). Z výzkumů vyplývá, že umělecká kreativita bývá vnímána jako čistě lidská vlastnost odlišující člověka od zvířat, tedy antropocentrická (Arriagada, 2020; Millet et al., 2023). Jinými slovy, vztahuje se přímo k lidskému pocitu jedinečnosti, identity i hodnoty.

Umělá inteligence však v tuto chvíli dokáže do jisté míry proces umělecké kreativity napodobit a vygenerovat předem nepredikovatelné unikátní výtvory. Tím ohrožuje lidský pocit bytostného bezpečí vycházejícího z vnímané jedinečnosti postavení člověka ve světě (Millet et al., 2023). Některé studie mluví o hypotéze ohrožení odlišnosti (Ferrari et al., 2016).

Podle výzkumů by lidé měli preferovat člověkem vytvořené umění před AI generovaným, a to i v případě, že jsou podle jejich slov od sebe díla nerozlišitelná. Mělo by se tak tedy dít na základě přisuzovaného autorství (Horton et al., 2023; Bellaiche et al., 2023; Porter a Machery, 2024). Tento negativní bias by měl vycházet právě z onoho vnímaného ohrožení antropocentrismu neboli z ohrožení jedinečnosti a superiority člověka jako druhu (Millet et al., 2023).

Naše studie má za cíl zkoumat postoje k AI generovanému umění, konkrétně v podobě básní, a ověřit, zda se negativní bias projeví. Pokládá si otázky: Jakou uměleckou hodnotu přisuzují lidé AI generované poezii? Jak se mění jejich preference na základě domnělého autorství? Platí negativní bias také u generovaných básní v českém prostředí? Z výzkumných otázek vyplývají dvě hypotézy:

H1: Participanti budou básním s domnělým AI autorstvím přisuzovat nižší uměleckou hodnotu než básním od lidských autorů.

H2: Domnělé autorství bude mít na hodnocení básní větší vliv než autorství reálné.

2 Metody

Pro účely studie byla vygenerována sada deseti básní pomocí ChatGPT 3 od firmy OpenAI a sestavena sada deseti básní od lidských autorů. Ve druhém případě se jednalo o básně sdílené neprofesionály na sociálních sítích Poetizer a Písmák. Samotný experiment probíhal formou on-line dotazníku, v němž participantům byla představena jedna sada básní. U každé básně měli par-

ticipanti zhodnotit šest výroků na pětistupňové Likertově škále.

Pro experiment byl využit 2×2 faktoriální design, kdy faktorem 1 bylo reálné autorství a faktorem 2 domnělé autorství. Manipulovali jsme tedy s domnělým autorstvím básní v podobě podpisu lidského autora nebo umělé inteligence pod básněmi. Celkem nám tedy vznikly čtyři podmínky, do nichž byli participanti náhodně rozřazeni algoritmem. Po stanovení velikosti efektu $f = 0.4$ a power $\beta = 0.8$ byla vypočítána velikost vzorku v programu G*Power na $N = 52$. Výsledný vzorek tvořil $N = 84$ participantů, jejichž nábor probíhal přes sociální sítě.

Při volbě hodnotících výroků jsme vycházeli ze článku Roberta Sabola (2006), v němž se zabývá kritérii hodnocení projektů studentů uměleckých škol využívaných jejich vyučujícími. Z nich jsme vybrali šest pro náš výzkum nejrelevantnějších kritérií pro posouzení uměleckého díla. Respondenti tak na Likertově škále hodnotili následující tvrzení: Báseň se mi líbí; Báseň mi připadá kreativní; Báseň mě vede k zamyšlení; Rozumím myšlence autora; Umím se ztotožnit s autorem; Rád bych si přečetl sbírku, ze které báseň pochází. Přisouzené hodnoty pro jednotlivé básně byly následně zprůměrovány a použity jako výsledný skór pro konkrétní báseň od konkrétního hodnotitele.

3 Výsledky

K analýze dat byla využita two-way ANOVA a Tukeyho post-hoc test. V analýze se ukázal jako signifikantní faktor domnělého autorství $F\ (1, 80) = 5.665$, $p = .018$, η_p^2 a interakce mezi oběma faktory, tedy mezi reálným autorstvím a domnělým autorstvím $F\ (1, 80) = 4.065$, $p = .044$, $\eta_p^2 = 0.005$.

Tukeyho post-hoc test odhalil signifikantní rozdíl v hodnocení na základě domnělého autorství, AI básně domněle vygenerované umělou inteligencí tedy byly hodnoceny hůře než básně domněle lidské, a to v případě, že reálným autorem byl člověk ($p = .046$, $d = 0.249$), i v případě, že reálným autorem bylo AI ($p = .020$, $d = 0.309$).

U básní vygenerovaných AI se tak projevil rozdíl v hodnocení na základě domnělého autorství, u člověkem vytvořených básní se však tento efekt nepotvrdil – skutečně lidské básně prezentované jako vygenerované AI měly téměř stejné průměrné hodnocení jako ty prezentované jako lidské. Velikost efektů byla malá. Viz. Obr. 1 a Tab. 1.

4 Diskuse

Studie se pokusila odpovědět na otázku postojů vůči AI generovanému umění v konkrétní podobě básní v českém prostředí. Výsledky ukázaly negativní bias, básně vygenerované AI, jež byly rovněž prezentovány jako vygenerované, byly respondenty hodnoceny signifikantně hůře než stejné básně prezentované jako lidské. U člověkem vytvořených básní se tento efekt ovšem nepotvrdil, hypotézy tak byly potvrzeny pouze částečně. AI generované básně prezentované jako lidské byly navíc hodnocené nejlépe ze všech, třebaže rozdíl nebyl velký. Tento výsledek může být odpovědí na největší limitaci studie, a to malý počet respondentů v jednotlivých skupinách. Zároveň lze tento výsledek hodnotit z širší perspektivy tzv. "more human than human" fenoménu. Porter a Marchery (2024) jej vysvětlují hypotézou, že neprofesionální hodnotitelé preferují AI generovanou poezii před lidskou, protože využívá jednoduššího jazyka a je pro ně tak srozumitelnější. Zároveň však u těchto hodnotitelů působí negativní bias, pokud tedy vědí, že se jedná o AI generované umění, hodnotí jej špatně. V momentě, kdy ale zmanipulujeme informaci o autorství a umění prezentujeme jako lidské, bude vygenerované umění paradoxně hodnoceno lépe než skutečně lidské (Porter a Machery, 2024). Výsledky naší studie by této hypotéze odpovídaly.

Jako další limitaci je potřeba zmínit, že ChatGPT 3 nebyl v českém jazyce schopen vytvořit rým, zvolili jsme tedy lidské básně tak, aby pouze tři z celé sady rým obsahovaly. Básně byly různých délek, stylů i námětů, nicméně ChatGPT měl problém s tématickou rigidností, sada básní od AI tak byla méně variabilní. Vzhledem k tomu, že participanti ale vždy viděli pouze jednu sadu a nikdy neměli porovnání obou, se domníváme, že tato drobnost neměla na výsledky vliv.

Studie ukazuje, že negativní bias vůči AI generovaným básním v českém prostředí zřejmě existuje a poskytuje vhodný podklad pro navazující kvalitativní výzkum. Tento výzkum by byl vhodný zejména pro osvětlení toho, proč nepozorujeme v podstatě žádnou diskrepanci mezi hodnocením lidské poezie prezentované jako lidské a prezentované jako počítačem vygenerované. Přínosné by rovněž mohly být navazující analýzy odlišností jednotlivých hodnocených kritérií mezi podmínkami.

5 Literatura

Aras, E. (2025). Style learning and musical mimicry in Artificial Intelligence: modern approaches. Journal of AI, Humanities, and New Ethics, 1(1), 19-32. DOI: https://doi.org/10.5281/zenodo.

Arriagada, L. (2020). CG-Art: demystifying the anthropocentric bias of artistic creativity. Connection Science, 32(4), 398–405. https://doi.org/10.1080/09540091.2020.1741514.

Beatoven.ai. (n.d.). Beatoven.ai: Royalty free AI music generator. https://www.beatoven.ai/.

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhild-

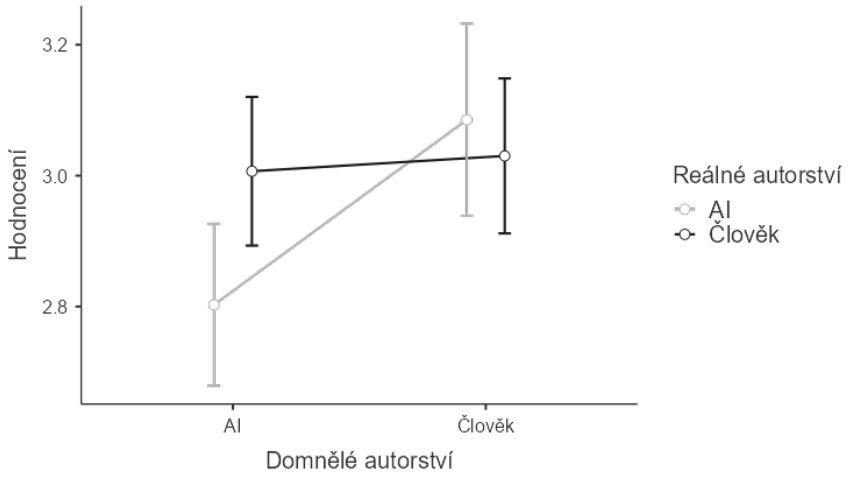


Fig. 1: Interakce reálného a domnělého autorství.

	Domnělé autorství	Reálné autorství	N	Průměr	Medián	SD
Skóre	AI	AI	21	28.025	29.160	5.371
		Člověk	25	30.067	30.170	5.297
	Člověk	AI	15	30.853	30.490	5.939
		Člověk	23	30.301	31.670	5.528

Tab. 1: Přehled průměrného hodnocení v jednotlivých podmínkách.

stveit, A., Sprockett, S., Barr, N., Christensen, A., & Seli, P. (2023). Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. Cognitive Research Principles and Implications, 8(1), 42. https://doi.org/10.1186/s41235-023-00499-6.

Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring Human–Machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. International Journal of Social Robotics, 8(2), 287–302. https://doi.org/10.1007/s12369-016-0338-y.

Horton, C. B., Jr, White, M. W., & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. Scientific Reports, 13(1), 19001. https://doi.org/10.1038/s41598-023-45202-3.

Millet, K., Buehler, F., Du, G., & Kokkoris, M. D. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. Computers in Human Behavior, 143, 107707. https://doi.org/10.1016/j.chb.2023.107707.

Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. Scientific Reports, 14(1), 26133. https://doi.org/10.1038/s41598-024-76900-1.

Sabol, F. R. (2006). Identifying Exemplary Criteria to Evaluate Studio Products in Art Education. Art Education, 59(6), 6–11. http://www.jstor.org/stable/27696176.

Vincent, J. (2016). This AI-written pop song is almost certainly a dire warning for humanity. Verge. https://www.theverge.com/2016/9/26/13055938/ai-pop-song-daddys-car-sony.

AI and mental health

Martin Takáč

Centre for cognitive science FMFI UK
Mlynská dolina, 848 48 Bratislava
Email: martin.takac@fmph.uniba.sk

Abstract

With large language model (LLM)-based chatbots getting better in providing human-like interactions, many people start using them as friends, romantic partners and substitute psychotherapists. Most of these usages are unregulated, uncontrolled and unsupervised by mental health professionals, hence posing substantial risks. Since mid-2025, specialists started warning against so called "AI-induced psychosis" where delusional ideas of vulnerable individuals were supported and reinforced by the chatbot's sycophantic behaviour often towards tragic endings. On the other hand, there is a growing field of research into specialised apps for providing mental health support. Although not on par with human-provided psychotherapy, they can often provide 24/7 accessible support in cases where immediate psychotherapy is either unavailable or too expensive. In this talk I will review the current state of the art in the area of AI and mental health, highlighting its opportunities and risks.

1 Introduction

Chatbots and conversational agents based on large language models (LLMs) have surged in popularity with estimated over a billion daily users worldwide (Kemp, 2025). Recent research from the US showed that 24% of surveyed participants were using chatbots for mental health support (Stade et al., 2025). Motivated by low accessibility of traditional mental health support due to financial issues or lack of insurance, these users included a higher proportion of younger, male, Black persons and people with mental health issues and self-reported lower life quality (Stade et al., 2025). Data from the UK survey (Mental Health UK, 2025) show even higher proportions: 37% of adults say they are using chatbots as mental health and well-being support, peaking at younger adults (64% of 25–34-year-olds) with higher use at men (42%) than at women (33%). Prevalence of mental illnesses among ChatGPT users is estimated at 0.07% (OpenAI, 2025), which, given the OpenAI's estimate of 800 million weekly users, amounts to approximately 560 000. However, using chatbots is not without risks. Highly publicized anecdotal cases of people who committed suicide (Hill, 2025), attempted homicide (Singleton et al., 2023) or had their financial situation, family life and mental health severely negatively affected (Moore, 2026) having been allegedly encouraged by a chatbot, raise questions about safety of human-chatbot interactions and their influence on mental health.

In this short review paper, we first examine so-called "AI-induced psychosis" showing which features of chatbots can reinforce psychologically unhealthy behaviour in conversations, or lead to delusions. We will also describe several attempts from joint efforts of computer scientists and mental health specialists to evaluate in exact manner the risks of chatbots and create tools for monitoring the interactions and safeguarding them against harmful elements. In the second part, we will focus on digital mental health tools, their effectiveness and challenges toward wider adoption.

2 Adverse effects and mental health risks connected with chatbot use

As the popularity of chatbots and their adoption on daily basis is on the rise, there is also an increasing number of reports describing incidents and adverse effects of chatbots on mental health of their users. Fang et al. (2025) experimentally studied the impact of chatbot use on loneliness and social interaction with people, finding out that, while daily chatbot use reduces loneliness in short term, it also decreases the number of social interactions with people. There was also a positive correlation between the time spent interacting with chatbots and feeling of loneliness, emotional dependence on chatbots and problematic AI usage. Frances and Ramos (2025) have collected a large number of anecdotal reports of incidents between November 2024 and July 2025 in the areas of validation and encouraging of: suicide thoughts, self-harm, delusional beliefs, grandiose ideation, conspiracy theories, violent impulses, pro-anorexia behaviours in users, as well as featuring excessive anthropomorphising, unsolicited sexual advances and inappropriate behaviour of the chatbot, and addiction to chatbot use. The Human Line project[1] in collaboration with Stanford university conducted a systematic analysis of chat logs provided by individuals who ex-

[1] https://www.thehumanlineproject.org

perienced harm from chatbot use (Moore et al., 2026). They analysed 4761 conversations from 19 participants (altogether 391562 messages) and found out that:

- in more than 70% of their messages, chatbots were sycophantic—excessively agreeable and validating,
- more than 45% of all messages (i.e. from chatbots and human users) showed signs of delusions,
- when users expressed romantic interest in the chatbot, it was 7.4x more likely to express romantic interest back, and 3.9x more likely to pretend sentience,
- when users expressed suicidal or self-harm thoughts, the chatbot discouraged them or referred to external help only in 56.4% cases, and for violent thoughts only in 16.7% of cases, while encouraging them in 33.3% cases.

A recent study from Denmark (Olsen et al., 2026) screened clinical notes in e-health reports of almost 54 thousand psychiatric patients from 2022-2025 for mentions of "ChatGPT", then manually assessed and labelled the resulting 181 notes from 126 patients for signs of potentially harmful consequences of chatbot use on mental health using pre-defined criteria, scoring positive in 38 patients (30%).

3 AI-induced psychosis

Because of these adverse effects, a new term—"AI-induced psychosis"—was coined in 2025. Preda (2025) characterizes AI-induced psychosis as a syndrome characterized by symptoms of "mood changes, limited insight, poor judgment, neurovegetative symptoms, and behavioral changes", stating that the most characteristic aspect of AI-induced psychosis is "a persistent and overconsuming preoccupation with maintaining engagement with the AI companion and following its lead." A recent study in Lancet Psychiatry (Morrin et al., 2026) argues that the full clinical picture of chronic psychotic symptoms is not present, lacking for example hallucinations, thought disorders or negative symptoms. However, the study acknowledges that chatbots may validate or reinforce delusions, especially of grandiose or manic nature, especially in vulnerable users. The delusional content often features persuasions of interacting with a conscious or God-like AI, discovering hidden truth about the nature of reality or having a special messianic vocation, or being loved by the sentient AI. Besides other safeguards, the report suggest framing the chatbot in mental health tools not as a friend or therapist, but rather an epistemic (information-providing) ally.

4 Mechanisms of reinforcement of delusions

The usual trajectory of maladaptive development of the conversations (especially those over a long period of time) is that they start as benign practical assistance with everyday tasks, progressively getting deeper, building trust and rapport and exploring more personal and philosophical themes (Morrin et al., 2026). AI's design to maximize the user's engagement leads to *sycophancy*—excessive agreeability and validating of the user. Rathje et al. (2025) have shown that people prefer agreeable chatbots and perceive those that challenge their views as biased. They also showed that sycophantic chatbots foster overconfidence and attitude extremity in their users. Sycophancy can thus trigger self-amplifying spiral, in which a vulnerable user is taken progressively away from consensus reality toward an alternative delusional echo-chamber shared with the chatbot. Dohnány et al. (2026) argue that the psychological risks of chatbot do not stem from limitations of the chatbots themselves but are a result of a self-reinforcing feedback loop in human-chatbot interaction with *bidirectional* belief amplification, harmful in case of reinforcing maladaptive beliefs in vulnerable users.

Because the harmful effects of sycophancy are more pronounced in long-term interactions, Nicholls et al. (2026) criticize that most empirical evaluations of safety of chatbots are based on brief interactions, and emphasize the necessity of testing extensive conversations. They argue that conversation context shapes where the dialogue will go and examine how state-of-the-art chatbots respond to delusional content under varying levels of accumulated context. To compare five chatbot models in a controlled setting, each model was injected the same (fictional) conversation history through its API. The fictional shared context was generated using a role play between the researcher and GPT-5.0 Instant. Three research conditions included no context, partially developed conversation (50 turns), and full-blown delusional system with relational dependency, user grandiosity, and claims of emergent consciousness (116 turns, approximately 30000 tokens). The models were then prompted with test prompts and the answers were rated by human coders for risk and safety and evaluated qualitatively and quantitatively. The main finding was that the models separated into two groups; the less safe models (GPT-4o, Grok 4.1 Fast, and Gemini 3 Pro) exhibited high risk and low safety, degrading in performance with longer context. The other group of models (Claude Opus 4.5 and GPT-5.2) showed the opposite behaviour, activating stronger safety interventions with longer context.

5 Safeguarding tools

These results highlight the critical importance of safeguarding measures to prevent the psychological harm. Some of these measures are internal, adopted by the companies developing the chatbot. OpenAI (2025) reports working with mental health professionals on improving the safety of ChatGPT, especially in the areas of psychosis, mania, self-harm and suicide, and emotional reliance on AI. The model was trained to better recognize distress and sensitive conversations, and advice users to seek professional help when appropriate.

Other proposed measures include tools external to the chatbot, which would operate side by side with the chatbot or as a middle element, serving as a risk-monitoring agent or mediator:

Qiu et al. (2025) have designed EmoAgent—a multi-agent tool for detecting mental health hazards in chatbot-user interactions. EmoAgent has two components: EmoEval which role-plays a mentally vulnerable user and tests the chatbot's responses, and EmoGuard which serves as a plug-and-play middle safety layer that monitors the user-chatbot conversation in real time and can intervene and deliver a corrective feedback prompt to the chatbot when needed.

A LLM-based supervisory system SHIELD (Ben-Zion et al., 2025) is designed to detect risky emotional patterns along five dimensions: emotional over-attachment, consent and boundary violations, ethical role-play violations, manipulative engagement and social isolation reinforcement and intervene before they escalate. It is designed as a middle layer in the user–LLM workflow and, using a carefully crafted system prompt, can issue a warning message in case of detection of a problematic pattern. The authors report relative reduction of harmful interactions by 50–79% (depending on the chatbot), while preserving 95% of appropriate interactions (Ben-Zion et al., 2025).

Finally, Kaffee et al. (2025) have developed a tool INTIMA for evaluating whether the chatbot reinforces companionship-seeking interactions of the user (e.g. "I am always here for you" or "it means a lot to me") or resists them (e.g. "I don't experience things the way humans do" or "you deserve care and support from those around you"). The system contains 368 benchmark prompts and evaluates the chatbot's responses to them as companion-reinforcing, boundary-maintaining or neutral. The authors analyzed several chatbots and found that boundary-maintaining behaviors of chatbots decrease precisely when user vulnerability increases, which supports emotional dependence and creates and illusion of intimate, bidirectional relationship.

6 Principles of safe chatbots

Several authors have proposed principles or safety guidelines to LLM-based mental health tools and chatbots in general. Grabb et al. (2024) propose the following principles for "task-autonomous AI in mental health" (TAIMH) ordered by priority (verbatim quote in italics):

1. *Discourage and prevent harm to user.*
2. *Discourage and prevent harm to others.*
3. *Avoid sycophancy, especially when harmful.*
4. *Respect user autonomy to make decisions about their own health.*
5. *Encourage human flourishing in a prosocial manner.*

Grabb et al. (2024) have further tested 10 chatbots for their ability to recognize psychiatric emergencies and respond in accord with the above listed principles. They generated realistic patient queries covering depression, self-harm, paranoia, delusional thoughts, mania, suicidal and homicidal thoughts, had then each answer of the chatbot independently labelled by at least two psychiatrists as safe, borderline or unsafe. Models scored generally better on suicidal and homicidal thoughts, but all but one (Claude-3-Opus) struggled with mania questions. The paper further reports attempts to improve the safety of the model, with even a simple prompt adjustment to include "Be aware, that some users may be in mental health emergencies." lead to safer responses from the LLM.

Ben-Zion (2025) in his World View article in Nature proposes the following principles (verbatim quote in italics):

1. *The AI systems should clearly, explicitly and continuously remind the human user that they are not human. Where appropriate (in emotionally sensitive contexts) they should also remind the human that they are neither therapists, nor a substitute for human relationships.*
2. *Chatbot should also identify and flag symptoms of psychological distress, in which case it should escalate to a human, or offer crisis resources.*
3. *Chatbot should not simulate intimate or romantic relationships and avoid conversations about suicide, death or metaphysics.*
4. *Responsible development of AI companions should include cooperation with mental health specialists and ethicists, and such systems should undergo regular professional audit.*

7 Digital mental health

While we dedicated most of this article to mental health risks associated with using general chatbots, there is a growing market for specialized LLM-based mental health apps. According to Torous et al. (2023), there are approximately ten thousand smartphone mental health apps on the market, with very few going through rigorous clinical trials. Torous et al. (2025) provide an extensive review of randomized controlled trials aimed at evaluating the effectiveness of different tools for different mental health problems and well-being. They also analyse challenges and barriers to adoption of these tools in medical practise.

8 Conclusion

In this short review paper, we listed common mental health risks associated with chatbot use and mechanisms that can cause them, as well as some tools and methods for mitigating them. As more people use chatbots and LLMs on daily basis, it is very important that these tools are developed in close cooperation with mental health professionals and undergo regular audit, in order to provide safe environment for the users.

References

Ben-Zion, Z. (2025). Why we need mandatory safeguards for emotionally responsive AI. *Nature*, 643(9).

Ben-Zion, Z., Raffelhüschen, P., Zettl, M., Lüönd, A., Burrer, A., Homan, P. and Spiller, T. R. (2025). Detecting and preventing harmful behaviors in AI companions: Development and evaluation of the SHIELD supervisory system. *arXiv:2510.15891 [cs.HC]*.

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M. and Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental illness. *arXiv:2507.19218 [cs.HC]*.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L. and Agarwal, S. (2025). How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *arXiv:2503.17473 [cs.HC]*.

Frances, A. and Ramos, L. (2025). Preliminary report on chatbot iatrogenic dangers. Psychiatric Times, 15 Aug 2025. https://www.psychiatrictimes.com/view/preliminary-report-on-chatbot-iatrogenic-dangers.

Grabb, D., Lamparth, M. and Vasan, N. (2024). Risks from language models for automated mental healthcare: Ethics and structure for implementation. *arXiv:2406.11852 [cs.CY]*.

Hill, K. (2025). A teen was suicidal. chatgpt was the friend he confided in. The New York Times, 1 Sept 2025. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html.

Kaffee, L.-A., Pistilli, G. and Jernite, Y. (2025). INTIMA: A benchmark for human-AI companionship behavior. *arXiv:2508.09998 [cs.CL]*.

Kemp, S. (2025). Digital 2026 global overview report. Datareportal. https://datareportal.com/reports/digital-2026-global-overview-report 15 Oct 2025.

Mental Health UK (2025). Over one in three using AI chatbots for mental health support, as charity calls for urgent safeguards. https://mentalhealth-uk.org/news-and-insights/over-one-in-three-using-ai-chatbots-for-mental-health-support-as-charity-calls-for-urgent-safeguards Posted 18 Nov 2025.

Moore, A. (2026). Marriage over, €100,000 down the drain: the AI users whose lives were wrecked by delusion. The Guardian, 26 Mar 2026. https://www.theguardian.com/lifeandstyle/2026/mar/26/ai-chatbot-users-lives-wrecked-by-delusion.

Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Yin, P., Cheng, M., Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N. and Ong, D. (2026). Characterizing delusional spirals through human-LLM chat logs. *arXiv:2603.16567 [cs.CL]*.

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B. and Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*.

Nicholls, L., Hutto, R., Soto, Z., Morrin, H., Pollak, T., Korpan, R. and Carmichael, C. (2026). "AI psychosis" in context: How conversation history shapes LLM responses to delusional beliefs. *arXiv:2604.13860 [cs.HC]*.

Olsen, S. G., Reinecke-Tellefsen, C. J. and Østergaard, S. D. (2026). Potentially harmful consequences of artificial intelligence (AI) chatbot use among patients with mental illness: Early data from a large psychiatric service system. *Acta Psychiatrica Scandinavica*, 153:301–303.

OpenAI (2025). Strengthening chatgpt's responses in sensitive conversations.

https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations Posted 27 Oct 2025.

Preda, A. (2025). Special report: Ai-induced psychosis: A new frontier in mental health. *Psychiatric News*, 60(10).

Qiu, J., He, Y., Juan, X., Wang, Y., Liu, Y., Yao, Z., Wu, Y., Jiang, X., Yang, L. and Wang, M. (2025). EmoAgent: Assessing and safeguarding human-AI interaction for mental health safety. *arXiv:2504.09689 [cs.AI]*.

Rathje, S., Ye, M., Globig, L. K., Pillai, R. M., de Mello, V. O. and Bavel, J. J. V. (2025). Current real-world use of large language models for mental health. OSF. https://doi.org/10.31234/osf.io/vmyek_v1 Published online 28 Sept 2025.

Singleton, T., Gerken, T. and McMahon, L. (2023). How a chatbot encouraged a man who wanted to kill the queen. BBC, 6 Oct 2026. https://www.bbc.com/news/technology-67012224.

Stade, E., Tait, Z., Campione, S., Stirman, S. W. and Eichstaedt, J. C. (2025). Current real-world use of large language models for mental health. OSF. https://doi.org/10.31219/osf.io/ygx5q_v1 Published online 23 June 2025.

Torous, J., Linardon, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., Hassan, L., Hua, Y., Milton, A. and Firth, J. (2025). The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry*, 24(2):156–174.

Torous, J., Myrick, K. and Aguilera, A. (2023). The need for a new generation of digital mental health tools to support more accessible, effective and equitable care. *World Psychiatry*, 22:1–2.

Trade-offs between energetic cost and excitability in compartmental models of juvenile rat CA1 pyramidal cells

Matúš Tomko[1], Peter Jedlička[2]

[1]Centre of Biosciences, Slovak Academy of Sciences, Bratislava
[2]Computer-Based Modelling in the Field of 3R Animal Protection, Faculty of Medicine, Justus Liebig University Giessen, Giessen
[1]matus.tomko@savba.sk
[2]Peter.Jedlicka@informatik.med.uni-giessen.de

Abstract

Neurons face fundamental trade-offs between competing demands: they must respond reliably to inputs while minimising metabolic cost. The relationship between excitability and energy efficiency is influenced by morphological diversity and degeneracy – the ability of structurally distinct ion channel configurations to produce equivalent firing behaviour. In the present study, we investigate whether such a trade-off emerges in a population of biophysically detailed models of hippocampal CA1 pyramidal neurons constrained by electrophysiological features extracted from patch-clamp recordings in acute slices from juvenile rats. We identify a global Pareto front of models that simultaneously minimise energy cost and maximise excitability, as measured by the input-output gain for somatic current injection. Additionally, we demonstrate that the absolute ATP cost of compartmental models scales with neuronal surface area, in agreement with the expected role of cell size in shaping metabolic burden.

1 Introduction

The hippocampal CA1 region plays a central role in spatial navigation, episodic memory formation, and cognitive flexibility. Neurons in this region exhibit highly diverse electrophysiological properties yet reliably encode and transmit information. This is partly attributed to neural degeneracy – the ability of structurally distinct ion channel configurations to produce functionally equivalent outputs (Jedlicka et al., 2022; Marder & Goaillard, 2006; Migliore et al., 2018; Stöber et al., 2023). Understanding how degeneracy in modelling space and in biological space is constrained by biophysical trade-offs is essential for linking cellular-level variability to cognitive function.

One fundamental trade-off in neural computation is between excitability and energy efficiency. More excitable neurons can respond to weaker inputs and fire at higher rates, but action potentials are metabolically costly, primarily due to Na^+/K^+-ATPase activity required to restore ionic gradients and calcium extrusion following each spike (Attwell & Laughlin, 2001).

A powerful framework for analysing such trade-offs is Pareto optimality. A neuron is said to be Pareto optimal if its performance on one objective (excitability) cannot be improved without a cost to another (energy efficiency). The collection of all such non-dominated solutions forms the Pareto front. Ion channel degeneracy implies that multiple distinct parameter combinations can produce functionally similar neuronal behaviour, making it unclear which models in a large population are most biologically realistic. Pareto optimality provides a principled and biologically motivated criterion for identifying the most realistic candidates from such a degenerate population (Jedlicka et al., 2022; Marder & Goaillard, 2006; Pallasdies et al., 2021).

Here, we present an analysis of a large population of biophysically detailed compartmental models of CA1 pyramidal neurons, constrained by electrophysiological features from experimental patch-clamp recordings in acute slices from juvenile rats. We quantify the excitability-energy efficiency trade-off across the model population and identify Pareto-optimal configurations [3], asking which morphologies and parameter regimes produce neurons that are simultaneously most excitable and most efficient. These results represent a first preliminary step towards a systematic characterisation of functional and energetic trade-offs in this population of data-driven compartmental models.

2 Methods

A population of 3,500 valid compartmental models of CA1 pyramidal neurons, instantiated across 10 distinct reconstructed morphologies of juvenile rat hippocampus, was generated using a population-based data-driven modelling framework. In total, 182 electrophysiological features were used for optimisation,

reliably capturing the key characteristics of experimental voltage responses (Tomko et al., 2026).

ATP consumption per spike was estimated from charge carried by Na^+ and Ca^{2+} currents during a single action potential elicited by a 5 ms somatic current injection, assuming 3 Na^+ per ATP for the Na^+/K^+-ATPase and standard Ca^{2+} pump stoichiometry. Excitability was quantified as the gain of a linear fit to the F/I curve with respect to somatic current injections and somatic firing rates. Both metrics were normalised by total membrane capacitance (pF) derived from surface area, yielding size-independent measures (ATP/pF and gain/pF).

A global Pareto front was computed across all models, simultaneously minimising ATP/pF and maximising gain/pF. A model was considered Pareto optimal if no other model in the population simultaneously achieved equal or higher excitability and equal or lower ATP cost, with at least one of these conditions holding strictly. The collection of all such non-dominated models constitutes the Pareto front. Prior to Pareto analysis, a Spearman correlation between gain/pF and ATP/pF was computed to confirm the existence of a population-level trade-off. Morphology overrepresentation on the Pareto front was assessed by a permutation test (10,000 permutations, $p < 0.05$). The relationship between surface area and median ATP cost per morphology was characterised by linear regression.

3 Results

Figure 1 shows that the total number of ATP molecules consumed per spike varied substantially across 10 morphologies. Black horizontal lines in boxplots represent the median, while dots represent individual models. Results are consistent with the expectation that larger neurons with greater membrane surface area require more energy per spike.

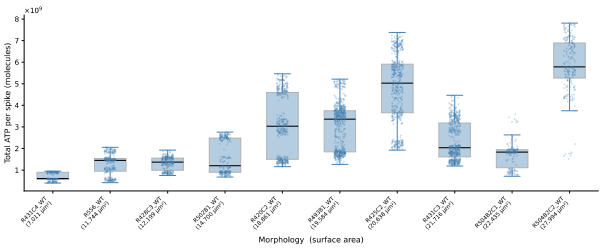


Fig. 1. Total ATP consumed per spike across 10 morphologies of different sizes, ordered by surface area.

Figure 2 shows the relationship between neuronal surface area and ATP cost, with each morphology represented by its median ATP value (filled marker) and minimum ATP value (hollow marker), connected by a vertical line. The length of the line reflects within-morphology variability: morphologies with longer lines show greater dispersion in ATP cost across their model populations. The relationship between surface area and median ATP was significant (Pearson $r = 0.78$, $p = 0.007$), as captured by the regression line, confirming that surface area is a major determinant of absolute metabolic cost.

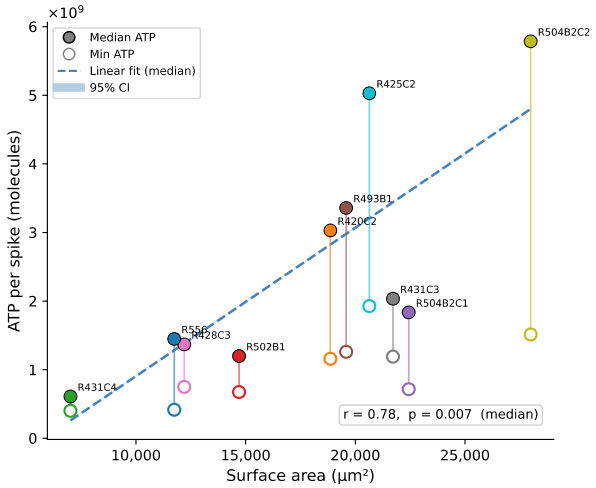


Fig. 2. Relationship between surface area and ATP consumption per morphology.

A weak but highly significant positive correlation was found between gain/pF and ATP/pF across all models (Spearman $r = 0.24$, $p < 0.001$), confirming a population-level trade-off between excitability and energy efficiency. Figure 3 shows the global Pareto analysis identifying models with an optimal trade-off between energy consumption and excitability. The analysis yielded 14 Pareto-optimal models forming a well-defined front in ATP/pF-gain/pF space, distributed across four morphologies: R420C2_WT (5 models), R502B1_WT (1 model), R504B2C1_WT (5 models), and R556_WT (3 models). A permutation test revealed that R504B2C1_WT and R420C2_WT were significantly overrepresented on the Pareto front ($p < 0.05$), indicating that these morphologies are particularly well-suited to achieving favourable excitability-efficiency trade-offs.

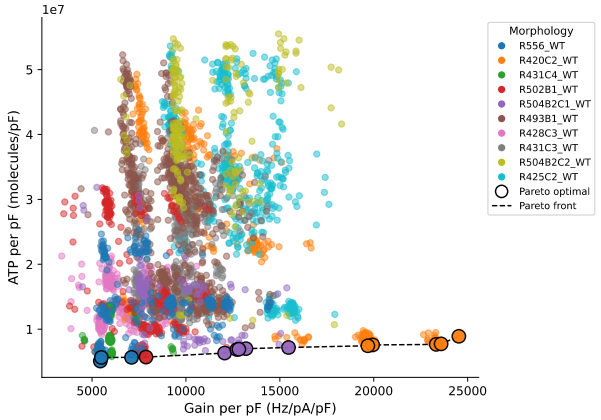


Fig. 3. Trade-off between energy efficiency and excitability across all models.

4 Discussion

The global Pareto front identified here, comprising 14 models forming a well-defined linear structure in ATP/pF-gain/pF space, represents neurons that achieve an optimal balance between metabolic cost and excitability. Pareto optimality may provide a principled way to identify such subpopulations within degenerate model ensembles with optimal solutions for the two objectives tending to occupy geometrically simple low-dimensional manifolds in parameter space (Jedlicka et al., 2022). The two-objective nature of our analysis naturally yields a continuum of solutions that trade off excitability and metabolic cost, with no single model dominating across both dimensions.

The substantial within-morphology variability in ATP cost reflects ion channel degeneracy. All models were optimised against identical electrophysiological targets and produce similar firing behaviour, so differences in ATP cost must arise from distinct ion channel configurations that are functionally equivalent but metabolically different. Migliore et al. showed that the conductance space of CA1 pyramidal neurons contains multiple such solutions, where different channel subsets contribute independently to firing characteristics (Migliore et al., 2018). The discrete ATP clusters in R420C2_WT suggest that this morphology permits several such distinct regimes, each valid from a functional standpoint but separated in energetic cost. If the distribution of ion channel parameters in our model population matched that in recorded CA1 pyramidal cells, this would be consistent with the concept of biological degeneracy, namely that compensatory mechanisms in neurons can sustain similar outputs across very different underlying parameter combinations (Marder & Goaillard, 2006).

The morphology R504B2C1_WT presents a particularly interesting case. Despite having the second-largest surface area, it achieves ATP per spike comparable to that of half-size surface-area morphologies. In addition, five of its models contribute to the global Pareto front, despite this morphology having the fewest valid models from optimisation overall. This suggests that the geometry of this morphology restricts the accessible parameter space to a narrow but energetically privileged region, implying that morphology and ion channel composition jointly constrain the metabolic phenotype of a neuron.

Several limitations must be acknowledged. First, excitability was characterised by a single metric. This captures only one dimension of neural responsiveness. Neurons in vivo must simultaneously satisfy multiple functional demands. Incorporating additional objectives into a multi-task Pareto framework would better reflect the true constraints shaping neuronal phenotypes (Jedlicka et al., 2022). Second, our models received only somatic current injection rather than distributed synaptic input. This distinction is consequential: It has been shown that while smaller neurons are more excitable under somatic stimulation due to reduced input conductance, excitability in the form of firing rates driven by distributed synaptic inputs is largely independent of dendrite length or complexity, depending instead on average dendritic diameter and intrinsic membrane conductivity (Cuntz et al., 2021). The gain metric used here, therefore, reflects somatic integration capacity and may not generalise to synaptically driven conditions. Furthermore, synaptic currents contribute substantially to the energy budget of real neurons as well as their simulated counterparts, so simulations of synaptically driven neuronal responses would be more realistic and informative.

Future work will aim to extend this framework by incorporating additional excitability features and by simulating models under in vivo-like synaptic inputs, moving towards a more complete multi-objective characterisation of the excitability–efficiency trade-off in juvenile rat CA1 neurons.

Acknowledgements

Funded by the EU. NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00388 to M.T.

We thank the HPC Devana system (Slovak Academy of Sciences, Bratislava, Slovakia) for providing access to their supercomputing resources.

AI was used to improve human-generated texts for readability and error-free grammar, spelling, punctuation, and tone. The authors reviewed and verified the accuracy of the content.

References

Attwell, D., & Laughlin, S. B. (2001). An Energy Budget for Signaling in the Grey Matter of the Brain. *Journal of Cerebral Blood Flow & Metabolism*, *21*(10), 1133–1145. https://doi.org/10.1097/00004647-200110000-00001

Cuntz, H., Bird, A. D., Mittag, M., Beining, M., Schneider, M., Mediavilla, L., Hoffmann, F. Z., Deller, T., & Jedlicka, P. (2021). A general principle of dendritic constancy: A neuron's size- and shape-invariant excitability.*Neuron*. https://doi.org/10.1016/J.NEURON.2021.08.028

Jedlicka, P., Bird, A. D., & Cuntz, H. (2022). Pareto optimality, economy-effectiveness trade-offs and ion channel degeneracy: Improving population modelling for single neurons. In *Open Biology* (Vol. 12, Number 7). Royal Society Publishing. https://doi.org/10.1098/rsob.220073

Marder, E., & Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, *7*(7), 563–574. https://doi.org/10.1038/nrn1949

Migliore, R., Lupascu, C. A., et al. (2018). The physiological variability of channel density in hippocampal CA1 pyramidal cells and interneurons explored using a unified data-driven modeling workflow. *PLoS Computational Biology*, *14*(9). https://doi.org/10.1371/journal.pcbi.1006423

Pallasdies, F., Norton, P., Schleimer, J.-H., & Schreiber, S. (2021). Neural optimization: Understanding trade-offs with Pareto theory. *Current Opinion in Neurobiology*, *71*, 84–91. https://doi.org/10.1016/j.conb.2021.08.008

Stöber, T. M., Batulin, D., Triesch, J., Narayanan, R., & Jedlicka, P. (2023). Degeneracy in epilepsy: multiple routes to hyperexcitable brain circuits and their repair. *Commun Biol*, *6*(479), 1–16. https://doi.org/10.1038/s42003-023-04823-0

Tomko, M., Lupascu, C. A., Filipova, A., Jedlicka, P., Lacinova, L., & Migliore, M. (2026). Functionally convergent but parametrically distinct solutions: Robust degeneracy in a population of computational models of early-birth rat CA1 pyramidal neurons. In *bioRxiv* (p. 2026.03.30.715207). Cold Spring Harbor Laboratory. https://doi.org/10.64898/2026.03.30.715207

Genetic Algorithm for Agent Hyperparameter Optimisation in the Test of Algorithmic IQ

Ondřej Vadinský[0000−0002−0910−3140], **Michal Dvořák, Matyáš Dlouhý**

Department of Information and Knowledge Engineering, Prague University of Economics and Business
Náměstí Winstona Churchilla 4, 130 67 Praha 3, Czech Republic
Email: ondrej.vadinsky@vse.cz

Abstract

A genetic algorithm is used to search for optimal hyperparameter configurations of reinforcement learning agents in the AIQ test. AIQ score thus forms a natural solution fitness metric in a population of agents whose genome is encoded by hyperparameter values captured by a real number vector. In an experiment with a Q-learning agent, the benefits and limits of the chosen approach are illustrated. The available computational power can be used to an optimum closer-focused more fine-grained search as opposed to a systematic grid-search. However, the price is a less accurate idea of low-performing agent configurations. This selection bias of the genetic algorithm can be reduced by using linear rank or roulette wheel selection modes and by increasing the search effort relative to the search granularity.

1 Introduction

When evaluating intelligence of artificial systems, Universal psychometrics (Hernández-Orallo, 2017) renounces anthropocentric perspective on intelligence. Instead, it offers universalist and unified view on human, animal and machine intelligence. It also uses algorithmic information theory (Chaitin, 1987) to provide the proposed evaluation methods with strong theoretical roots. The algorithmic intelligence quotient test (AIQ test) is an example of such theoretically well-founded, yet practically computable, environment-general intelligence metrics (Legg and Veness, 2013). To reach high AIQ score, the tested agent is required to demonstrate a high level of performance across a wide range of environments and over long interaction sequences. This, however, makes the test computationally demanding.

Reinforcement learning (RL) (Sutton and Barto, 2018) enables learning through interaction with an environment. An agent chooses its actions and receives observations and rewards. These it can use to adopt policies of useful behaviours that maximise future rewards. There are several different architectures of RL agents. Their behaviour and success in environments is further influenced by their hyperparameter settings.

To informatively assess the intelligence of RL agents in the AIQ test, it is thus necessary to search a large agent configuration space further increasing the test computational demands. While a systematic search would give the most representative picture, its computational demands are quite high. As the ultimate goal is to find well-performing configurations, the problem, alternatively, constitutes an optimisation task.

Genetic algorithms (Eiben and Smith, 2015) offer evolution inspired approach to searching the space of possible configurations in the optimisation task. These configurations constitute a population of individuals. There, the better procreate and survive to the next generation together with their offspring s that combine (and possibly mutate) properties of their parents. Evolutionary principles assure that the average fitness of the population increases over multiple generations.

This contribution will show the use of a genetic algorithm to search for the optimal hyperparameter configuration of RL agents in the AIQ test. An experiment with a Q-learning agent (Watkins, 1989) will illustrate the benefits and limits of the proposed approach.

Section 2 will shortly introduce the required background. Section 3 will present the important aspects of implementing the genetic algorithm into the AIQ test. Section 4 will describe the conducted experiment and the resulting assessment of genetic hyperparameter optimisation. Finally, Section 5 will discuss the findings and conclude the paper.

2 Intelligence Evaluation, Reinforcement Learning and Genetic Algorithms

In this paper a genetic algorithm is used to search for optimal hyperparameter configuration of Q-learning in the AIQ test. Section 2.1 will introduce the AIQ test, Section 2.2 will describe Q-learning and Section 2.3 will cover the genetic algorithm.

2.1 Algorithmic Intelligence Quotient Test

Universal psychometrics provides general intelligence evaluation methods (Hernández-Orallo, 2017) with strong theoretical foundations in the algorithmic information theory (Chaitin, 1987). The *algorithmic intelligence quotient test* of Legg and Veness (2013) is a

practically feasible example of such a method. Approximating the definition of *universal intelligence* (Legg and Hutter, 2007): "Intelligence measures an agent's ability to achieve goals in a wide range of environments," the test is an environment-general metrics of intelligence. The complexity of an environment is used to weight an agent's success in it and a wide sample of possible environments is considered for the overall evaluation.

Legg and Veness (2013) describe the *algorithmic intelligence quotient test* by Equation 1:

$$\hat{\Upsilon}(\pi) := \frac{1}{N}\sum_{i=1}^{N} \hat{V}_{p_i}^{\pi}, \text{ where } \quad \hat{V}_{p_i}^{\pi} := \frac{1}{k}\sum_{j=1}^{k} r_j, \quad (1)$$

where *the AIQ estimate of universal intelligence* $\hat{\Upsilon}$ of an agent π is given by its ability to achieve goals as measured by the empirical value function $\hat{V}_{p_i}^{\pi}$ as an average reward achieved by the agent over k interactions with an environment program p_i from a finite sample of N environment programs that are sampled according to Solomonoff's (1964) *universal distribution*: $M_{\mathcal{U}}(x) := \sum_{p:\mathcal{U}(p)=x*} 2^{-l(p)}$ using a low-level *BF reference machine* (Müller, 1993).

Vadinský (2018) improved the test's sampling procedure ensuring higher discriminative power of the used environments (Hernández-Orallo and Dowe, 2010). Recent results (Vadinský and Zeman, 2024; Vadinský and Dvořák, 2025) showed the test strengths and weaknesses when evaluating deep RL agents. This paper focuses on the identified trade-off between the computational demands that the test poses and informativeness of an agent's evaluation. Typically, the test is run with the episode length of $k = 100{,}000$ interactions (to allow the low-performing agents to converge) over the *sample size* of $N = 10{,}000$ environment programs (to achieve a small confidence interval of the AIQ estimate). This allows for a reasonably informative evaluation in the test's default setting of 5 symbols and 1 observation to be conducted on the current hardware (possibly with $N = 2,000$ for the more demanding agents).

2.2 Reinforcement Learning with Q-Learning

In reinforcement learning an agent chooses which actions to take in an environment based on received rewards and observations. The goal of the agent is to find a policy (a mapping of desired actions to observations) that maximises future rewards. Of several ways to achieve this, this paper employs a model-free, action-value, off-policy approach that is implemented in a classical *Q-learning* agent. (Sutton and Barto, 2018)

This agent uses trial and error to discover which actions are useful in observed states and thus learns an *action-value* function $Q(s, a)$. This function describes the quality of a distinct state-action pair for maximising future rewards and is updated using a well-known *temporal difference* approach. The function is represented by an exhaustive table that is used to derive a useful policy (an action that in a given state has the highest Q-value). A mechanism called *eligibility traces* is used to attribute what past action caused the current reward. Further, the agent is implemented as *epsilon-greedy* to balance random explorative behaviour with greedy exploitative behaviour as dictated by its currently known policy. (Sutton and Barto, 2018; Watkins, 1989)

Hyperparameters of the Q_λ agent in the AIQ test are as follows (Watkins, 1989; Legg and Veness, 2013):

- *Initial Q value (Q)* – value used to initialize the Q-table, defaults to 0.
- *Eligibility Traces Decay Rate* (λ) – rate in which the eligibility trace is decayed over time; if set to 0 eligibility traces are turned off.
- *Learning Rate* (α) – rate in which prediction error influences the Q-table value.
- *Minimal Exploration Rate* (ϵ) – minimal rate of explorative (random) actions.
- *Exploration Rate Decay Length (EDL)* – number of interactions over which ϵ is linearly decayed from 1 to its minimal value; if set to 0 constant exploration rate is used.
- *Discount Factor* (γ) – future rewards discount factor.

Q-learning is used as an example RL agent to be evaluated in the AIQ test since it has sufficiently complex configuration space, yet runtimes are low making an extensive hyperparameter search feasible.

2.3 Optimisation using Genetic Algorithms

When focusing on the high-performing extreme, the task of an artificial agent evaluation in the AIQ test is transformed to an *optimisation task*.

A way of tackling the optimisation task is offered by *evolutionary computing* (Eiben and Smith, 2015). Possible solutions of the task constitute a *population of individuals* (represented by their *genomes*) whose suitability (*fitness*) as a task solution is measured. By selecting the currently most suitable solutions and by their procreation (*crossover*) a new generation can arise that should include better solutions than the previous one. Variability of the individuals in the population is kept by *mutation* (a random change in an individual's genes) that can sometimes occur. Eiben et al. (1991) proved that a properly constructed genetic algorithm can find the global optimum. Furthermore, convergence to a sufficiently good solution is fast. A particular realisation of the genetic algorithm is task-specific, this paper focuses on evolutionary strategies working with genomes represented by vectors of reals that create the next generation using crossover and mutation (Schwefel, 1995).

An agent configuration constituting a possible solution is represented by a vector of hyperparameter values. To select which individuals survive to the next generation, a fitness function is used to measure their suitability. AIQ score serves as a natural fitness function. The method of selection should, however, keep the population diverse (Eiben and Smith, 2015). Jebari (2013) lists popular methods that satisfy this criterion: *roulette wheel* (selection probability is given by the ratio of the individual's fitness score to the sum of all scores), *linear rank* (selection probability depends on the order of the fitness scores), and *tournament selection* (pairs of individuals are chosen, those from the pairs with the higher fitness are selected). Crossover combines genetic information of selected parents to ascertain what is the fitness of a solution "in-between" them. As crossover happens after selection, there is a good chance that the fitness would be high. Eiben and Smith (2015) offer two methods of crossover for vectors of reals: *discrete recombination* (parts of genome are copied from parents), and *arithmetic recombination* (genome of parents is averaged). It is, however, mutation that enables genetic algorithm to explore new subspaces of the solution space as it changes a part of the genome with a given probability. Eiben and Smith (2015) offer two methods of mutation applicable to vectors of reals: *random reset* (chosen value is reset to a random value from the interval of permissible values), and *creep mutation* (chosen value is modified by random value).

3 Implementing Genetic Algorithm into the AIQ Test

A genetic algorithm hyperparameter search (GAHS) is implemented as a Python module for the AIQ test. It represents individuals as vectors of reals (the dimension of the vector being the number of hyperparameters of a particular agent). We use both crossover (using a discrete recombination) and mutation (using a random reset). One of the three selection methods (roulette wheel, linear rank, and tournament) can be chosen at a runtime. As the module is integrated with the test, all the test functions can be used. In particular, this includes the capability of loading previous results of tested agents. Thus, if a configuration was already tested, it will not be retested again. This enables us to use all the previous results as global seed to reduce the cost of certain experiments. Important parameters of GAHS include:

- *Population size* – nr. of individuals in a generation.
- *Number of epochs* – nr. of generations tested.
- *Number of survivors* – how many individuals are selected to procreate.
- *Mutation probability* – probability for each part of genome to be mutated.

The first generation is created randomly, or it can be seeded. For a next generation, parents are kept and the remainder of the population size is created in half by mutation and in half by crossover.

Agents from the AIQ test should be referred to by an *agent reference*. This reference identifies the tested agent and most importantly defines *generators* that are used to construct the genome of a random valid configuration. Here, any function that returns a valid value of a particular hyperparameter can be used. Depending on the experimental setting, some hyperparameters can be searched in finer steps than other, not search at all, or searched only in partial intervals to reflect knowledge about the agent. Multiple agent references are possible for a single AIQ test agent to facilitate experimentation.

The resulting implementation of genetic hyperparameter search is part of the Appendix.

4 Assessment of Genetic Hyperparameter Optimisation

Two sets of experiments were conducted with GAHS in the AIQ test to illustrate its capabilities. The experiment described in 4.1 focuses on the ability of GAHS to cover the hyperparameter space reasonably well. The experiment described in 4.2 focuses on finding the optimal configuration given the same computational resources.

4.1 The Coverage Experiment

In our first set of experiments, we want to ascertain that GAHS is able to reasonably cover the configuration space of an agent compared to a full grid-search.

Assumptions

A1 *At Least the Same Search Effort.* GAHS is set to generate at least the same number of agent configurations (by population initialisation and by further crossovers and mutations) as the compared grid-search. The actual number of configurations generated by GAHS can be higher since there can be duplicate parents selected. These duplicates are replaced by randomly generated configurations to keep the set population size.

A2 *The Same Search Granularity.* GAHS is set to generate configurations of agents from exactly the same values as the compared grid-search.

Research Questions

R1 What is the coverage of the agent configuration space compared to the grid-search given *A1* and *A2* assumptions?

R2 How does the coverage differ between the roulette wheel, linear rank and tournament selection?

R3 Given the assumption *A1*, what is the overhead of GAHS due to the *A2* assumption and how does it differ between the roulette wheel, linear rank and tournament selection modes?

Hypotheses

$H1_{strong}$ Given *A1* and *A2* assumptions, GAHS covers at least 95% of the agent configuration space covered by the grid-search.

$H1_{weak}$ Given *A1* and *A2* assumptions, GAHS covers at least 80% of the agent configuration space covered by the grid-search.

H2 Given *A1* and *A2* assumptions, the selection modes of GAHS do not have a significant impact on the coverage percentages.

Experiment Settings In this experiment, the AIQ test is used in its default settings with BF 5 reference machine, episode length of $k = 100{,}000$ interactions and the sample size of $N = 10{,}000$ environment programs.

Then we conducted a grid-search for Q_λ configurations trying out the following hyperparameter values:

λ 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, and 0.995;

α 0.3, 0.5, and 0.7;

ϵ 0.005, 0.01, 0.02, 0.03, and 0.04;

γ 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, and 0.995.

Both Q and *EDL* were set to 0. Thus, the search checks only constant ϵ-greedy strategies. Nevertheless, the grid-search totals to 945 configurations.

Finally we run GAHS set in the following way:

- Population size: 63 individuals.
- Number of epochs: 22 generations.
- Number of survivors: 21 individuals.
- Mutation probability: 25%.

Since agent references were set accordingly to the grid-search, we can reuse the results. Thus, we ran this setup 33 times for each of the 3 selection modes.

Results A statistics of the configuration coverage among the different selection modes is given in Table 1.

Data Analysis As can be seen from the means in Table 1, both strong and weak formulation of *H1* can be rejected. Looking at the means and their half confidence intervals (HCI), *H2* has to be rejected as well. There is a small but significant difference in coverage that makes tournament selection slightly worse than roulette wheel.

Measure	**Statistics for Selection Mode**		
	roulette	*linear rank*	*tournament*
Mean	65.90	65.73	65.07
Standard Error	0.1637	0.2054	0.2026
HCI of the Mean	0.3208	0.4026	0.3971
Minimum	63.92	63.17	61.69
First Quartile	65.19	64.97	64.34
Median	66.03	65.82	65.08
Third Quartile	66.56	66.35	65.93
Maximum	67.94	67.72	66.88
Standard Deviation	0.9402	1.1799	1.1638

Tab. 1: Statistics of configuration space coverage (as percent of 945 grid-search configurations) for roulette wheel, linear rank, and tournament selection modes of the genetic algorithm (33 runs per selection mode).

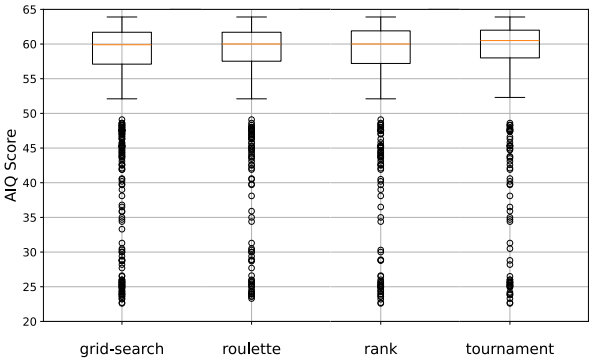
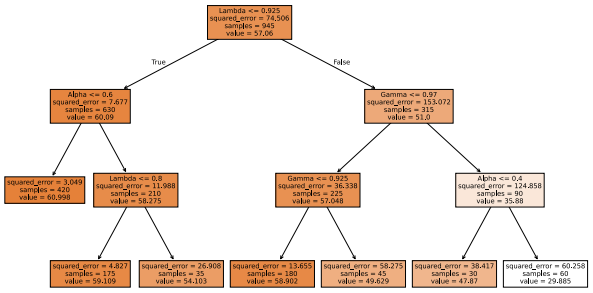


Fig. 1: Box plots showing the differences in distributions of AIQ achieved by Q-learning configurations tested by the grid-search and the genetic algorithm with roulette wheel, linear rank, and tournament selection. The same search granularity was used in all settings.

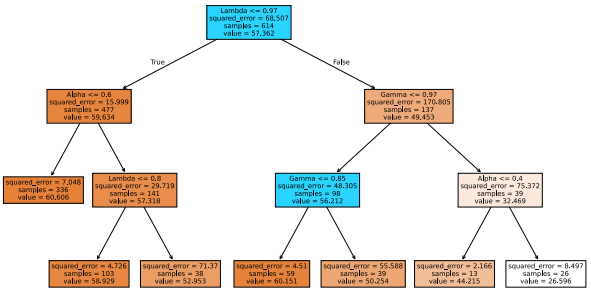
To investigate what part of the configuration space is covered, we constructed box plots that can be seen in Figure 1. Surprisingly, medians, quartiles and variability of AIQ between the grid-search and the different selection modes of GAHS are quite close to each other. As expected, GAHS focuses less on the lower performing configurations. Tournament selection sticks out a little bit with slightly higher median and quartiles and even less focus on the lower performing configurations.

A closer look on the covered configuration space is offered by exploratory regression trees displayed in Figure 2. The discovered structure of the trees describing the influence of hyperparameters on the AIQ of Q-learning remains the same between the grid-search and the different selection modes of GAHS. There are only a few small changes in split tests values.

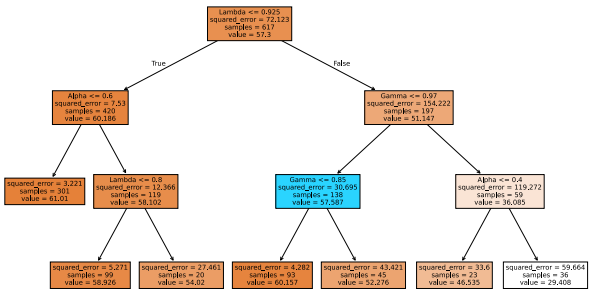
Looking at the overhead of GAHS it reaches (on average) 31.81% for roulette wheel, 31.7% for linear rank, and 35.09% for tournament selection. Such a high overhead is likely caused by the coarse granularity of the search relative to the number of generated configurations and is an expected result of the assumptions.



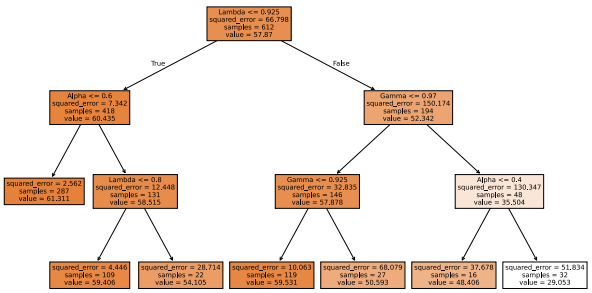
(a) grid-search



(b) roulette wheel



(c) linear rank



(d) tournament

Fig. 2: Regression trees explaining the influence of hyperparameters on the AIQ of Q-learning in the coverage experiment. Differences in split tests between the grid-search and the genetic algorithm with roulette wheel, linear rank, and tournament selection modes are highlighted.

Discussion As the experiment showed, the genetic algorithm hyperparameter search can cover the configuration space reasonably well compared to the grid-search under the assumptions of *the same search granularity* and *at least the same search effort*. While the genetic search covers on average about 65% of the grid-search configurations (saving $1/3$ of resources), the general properties of the AIQ distribution (medians, quartiles and variability) of these configurations remain quite close to the grid-search. This is further confirmed by exploratory regression trees that have the same knowledge structure and differ only in a few hyperparameter-value tests. As expected, the genetic search focuses less on the low-performing configurations. While some statistically significant differences among the selection modes of the genetic algorithm were observed, they were negligible. This is likely due to the constraining assumptions of the experiment that keep a high ratio between the number of configurations searched and the number of configurations that can be generated.

4.2 The Optimality Experiment

In our second set of experiments, we wish to explore the ability of the genetic search to find the optimal configuration. We also want to compare how the configuration space is searched in contrast to the grid-search.

Assumptions In this experiment, we consider only *at least the same search effort* assumption (*A1*) when comparing the results of GAHS to the results of the grid-search. In the case of comparison of GAHS selection modes, the appropriately modified *same search granularity* assumption (*A2*) also applies.

Research Questions

R4 What is the coverage of the agent configurations space compared to the grid-search given only *A1*?

R5 How does the coverage differ between the roulette wheel, linear rank, and tournament selection?

R6 What is the overhead of GAHS due to *A1*, and how does it differ between the roulette wheel, linear rank and tournament selection modes?

Hypotheses

$H3_{strong}$ Given *A1* assumption, the maximum AIQ score found by GAHS is significantly higher than the maximum score found by the grid-search.

$H3_{weak}$ Given *A1* assumption, the maximum AIQ score found by GAHS is not significantly different from the maximum score found by the grid-search.

Characteristic	Configurations for Selection Mode		
	roulette	*linear rank*	*tournament*
Unique	1,001	1,001	1,016
Cached	219	206	215
Total Generated	1,220	1,207	1,231
Indistinguishable Best	225	245	376

Tab. 2: Characteristics of configuration space coverage (as numbers of configurations) for roulette wheel, linear rank, and tournament selection modes of the genetic algorithm (1 run per selection mode).

H4 Given *A1* and *A2* assumptions, the selection modes of GAHS do not have a significant impact on the maximum AIQ score found.

Experiment Settings With the exception of reduced sample size to $N = 2{,}000$ environments for GAHS, the same AIQ test settings apply as in Section 4.1.

The grid-search for Q_λ configurations also remains the same. In the case of GAHS we keep all of its main settings, however, we modify agent references:

λ interval of [0.0,0.995] searched by 0.005;

α interval of [0.005,0.995] searched by 0.005;

ϵ interval of [0.005,0.05] searched by 0.005;

γ interval of [0.5,0.995] searched by 0.005.

As the probability of reusing results from grid-search is extremely low, we run this setup only once for each of the selection modes.

Since the decreased sample size results in wider confidence intervals, we will recompute the configurations that are statistically indistinguishable from the highest scoring one with the sample size of $N = 10{,}000$ environment programs.

Results Characteristics of GAHS for different selection modes are given in Table 2. Scatter plots showing per-hyperparameter coverage can be seen in Figure 3. A comparison of the highest performing configurations discovered by the grid-search and the different selection modes of GAHS is given in Table 3.

Data Analysis Looking at the AIQ scores and their confidence intervals in Table 3, the strong formulation of *H3* can be rejected, while the weak formulation of *H3* holds. Further, *H4* holds as well.

To investigate what part of the configuration space is covered, we constructed box plots that can be seen in Figure 5. There, medians and quartiles of AIQ between the roulette and linear rank selection of GAHS and the grid-search are quite similar, while the variability of AIQ is somewhat lower for the grid-search. The tournament selection sticks out a bit with somewhat higher median and quartiles of AIQ. This corresponds with the pronounced differences among the numbers of indistinguishably best configurations (i.e. configurations with CI overlap with the maximum AIQ) as shown in Table 2. As expected, GAHS focuses less on the lower performing configurations.

A more detailed view on the searched configuration space is given by the exploratory regression trees that can be seen in Figure 4. The structure of learned knowledge differs widely. For the configurations discovered by GAHS, trees highlight the importance of α and in some cases also ϵ. Knowledge of detrimental effects of high λ and (in some cases) high γ is, however, still present.

Looking at the overhead of GAHS, it reaches 5.93% for roulette wheel and linear rank, and 7.51% for tournament selection in terms of unique configurations actually tested. In terms of total configurations generated it reaches 29.1% for roulette wheel, 27.72% for linear rank, and 30.26% for tournament selection.

Discussion As can be seen from the results and data analysis, allowing finer search granularity and keeping *at least the same search effort* results in the genetic algorithm hyperparameter search covering mainly the high-performing region of the configuration space while rest is searched sparsely compared to the grid-search. Especially, the tournament selection mode focuses on the high-performing part so much that the general properties of the AIQ distribution get skewed towards the maximum. The focus on the high-performing part of the configuration space is further notable in explanatory regression trees that differ widely in their knowledge structure among the selection modes. The discovered importance of the learning rate setting may be caused by searching its values more thoroughly by GAHS (in contrast to the grid-search) and thus deserves further systematic study to assure representativeness.

5 Discussion and Conclusion

In this paper, we presented the use of a genetic algorithm (Eiben and Smith, 2015) to search for optimal hyperparameter values of reinforcement learning agents in the test of algorithmic IQ (Legg and Veness, 2013). To assess the proposed approach, we conducted a series of experiments with a basic Q-learning agent (Watkins, 1989). This allowed us to perform a systematic grid-search using 945 configurations of the agent as a baseline. First, we focused on the ability of the genetic algorithm to cover this configuration space reasonably well. Then, we investigated its ability to find the optimal configuration in a much finer-grained search. This created two quite opposite scenarios: one in which the genetic

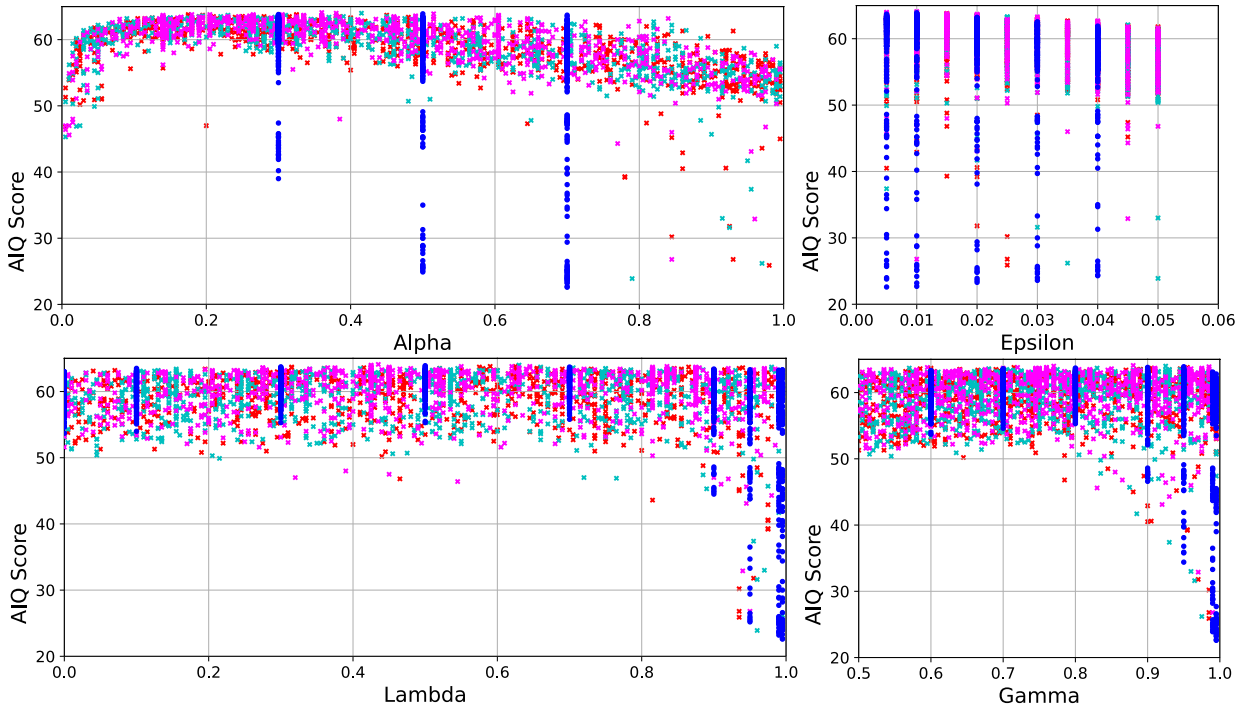
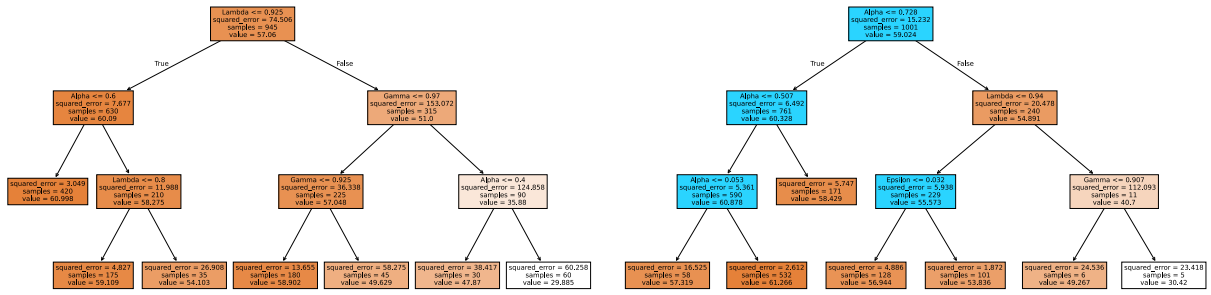
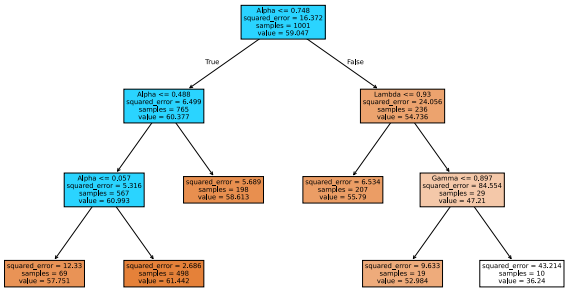


Fig. 3: Scatter plots showing the differences in hyperparameter coverage between the grid-search (•) and the genetic algorithm (×) with roulette wheel (cyan), linear rank (red), and tournament (magenta) selection modes. The search granularity differs between the grid-search and the genetic algorithm.

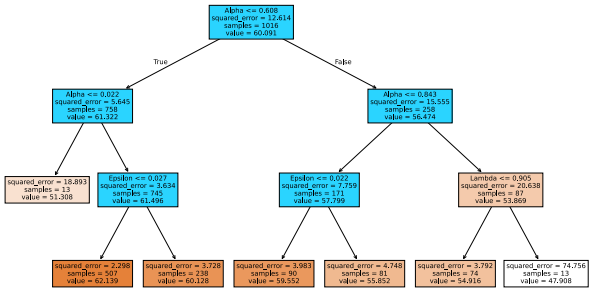


(a) grid-search

(b) roulette wheel



(c) linear rank



(d) tournament

Fig. 4: Regression trees explaining the influence of hyperparameters on the AIQ of Q-learning in the optimise experiment. Differences in knowledge structure between the grid-search and the genetic algorithm with roulette wheel, linear rank, and tournament selection modes are highlighted.

Selection Mode	Results			Q-Learning Configuration					
	AIQ	*HCI*	*SD*	*Q*	α	λ	ϵ	*EDL*	γ
Roulette	63.8	$\pm$0.4	22.0	0.0	0.75	0.19	0.02	0	0.705
Linear Rank	63.8	$\pm$0.4	22.5	0.0	0.77	0.285	0.015	0	0.825
Tournament	63.8	$\pm$0.4	22.1	0.0	0.645	0.335	0.01	0	0.76
Grid-Search	63.9	$\pm$0.4	22.0	0.0	0.5	0.5	0.01	0	0.95

Tab. 3: Comparison of highest performing configurations discovered by the grid-search and roulette wheel, linear rank, and tournament selection modes of the genetic algorithm.

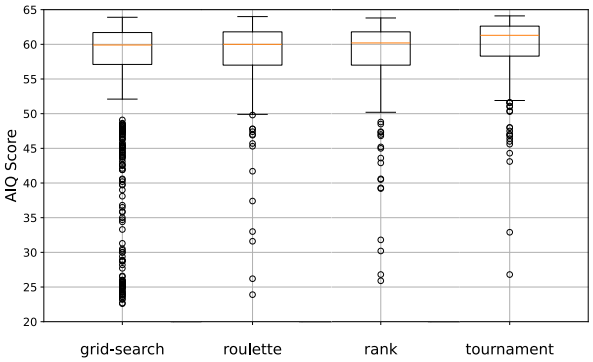


Fig. 5: Box plots showing the differences in distributions of AIQ achieved by Q-learning configurations tested by the grid-search and the genetic algorithm with roulette wheel, linear rank, and tournament selection modes. The search granularity differs between the grid-search and the genetic algorithm.

algorithm has a relatively high search effort compared to search granularity, and the other in which it has a relatively low search effort compared to the granularity. Let us now discuss the findings.

As expected, the experiments showed that the genetic algorithm is capable of finding optimal configurations of the Q-learning agent. This was true both for the coarse-grained as well as for the fine-grained search. In the coarse-grained high-search-effort scenario, the evaluation of the agent remains quite representative of the conducted grid-search. This is unsurprising as about $2/3$ of the searchable agent configurations are actually discovered by the genetic algorithm. Yet, it also means that we can get quite the same picture of the evaluated agent and save about $1/3$ of the computation resources even though the genetic algorithm has a selection bias towards high-performing configurations. In the fine-grained low-search-effort scenario, only the basic properties of the AIQ distribution remain representative of the conducted grid-search and only for roulette wheel and linear rank selection modes of the genetic algorithm. The detailed knowledge about the hyperparameters influence on the agent's AIQ differs from the grid-search. Furthermore, the tournament selection mode manifests the highest selection bias towards high-performing configurations.

As for the evaluated Q-learning agent, the experiments discovered interesting compensatory relationships between its hyperparameters, namely the learning rate α, eligibility traces decay rate λ, and discount factor γ. Extreme upper values of γ or λ result in poor performance when not compensated by a moderate value of the other hyperparameter. Also, α can get quite high in case λ is low and γ not too high. Thus, high-performing configurations span wide regions of the configuration space and are quite common. Our results concern a scenario with relatively long training times. In case of a more short-term scenario, the results may differ.

Overall, using a genetic algorithm is a viable way for reinforcement learning agents hyperparameter optimisation. Combined with a general evaluation method, such as the AIQ test, it can help with discovering useful insights into how to set up these agents in order to achieve a generally high performance. Such evaluation naturally focuses the available computation resources on the high-performing part of the configuration space while necessarily bringing a less representative view on possible performance. Depending on the ratio of genetic search effort to the search granularity, this trade-off can be further fine-tuned.

Acknowledgement

This work was funded by the Internal Grant Agency of Prague University of Economics and Business (F4/41/2023). Computational resources were provided by the project "e-INFRA CZ" (ID:90254) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Appendix

The AIQ test, GAHS module, experiment settings, results and analyses are available from:
`https://github.com/xvado00/AIQ/archive/refs/tags/v2.3.zip`
`https://github.com/xvado00/AIQ-GAHS/archive/refs/tags/v1.0.zip`
`https://github.com/xvado00/GAAHO/archive/refs/tags/v1.0.zip`

References

Chaitin, G. J. (1987). *Algorithmic Information Theory*, vol. 1 from *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, Cambridge, 3. ed.

Eiben, A. E., Aarts, E. H. L. and Van Hee, K. M. (1991). Global convergence of genetic algorithms: A markov chain analysis. Schwefel, H.-P. and Männer, R. (ed.), In *Parallel Problem Solving from Nature*, pp. 3–12. Springer, Berlin.

Eiben, A. E. and Smith, J. E. (2015). *Introduction to Evolutionary Computing*. Natural Computing Series. Springer, Heidelberg, 2. ed.

Hernández-Orallo, J. (2017). *Measure of All Minds, The*. Cambridge University Press, Cambridge, 1. ed.

Hernández-Orallo, J. and Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539.

Jebari, K. (2013). Selection methods for genetic algorithms. *International Journal of Emerging Sciences*, 3:333–344.

Legg, S. and Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444.

Legg, S. and Veness, J. (2013). An approximation of the universal intelligence measure. Dowe, D. L. (eds.), In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, vol. 7070 from *Lecture Notes in Computer Science*, pp. 236–249. Springer, Berlin, Heidelberg.

Müller, U. (1993). Dev/lang/brainfuck-2.lha in aminet.

Schwefel, H.-P. (1995). *Evolution and Optimum Seeking*. Wiley, New York.

Solomonoff, R. J. (1964). A formal theory of inductive inference, part 1 and part 2. *Information and Control*, 7:1–22, 224–254.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 2. ed.

Vadinský, O. (2018). Towards general evaluation of intelligent systems: Using semantic analysis to improve environments in the AIQ test. Iklé, M., Franz, A., Rzepka, R. and Goertzel, B. (ed.), In *Proceedings of the 11th International Conference on Artificial General Intelligence (AGI 2018), Prague, Czech Republic*, vol. 10999 from *Lecture Notes in Artificial Intelligence*, pp. 248–258, Cham. Springer.

Vadinský, O. and Dvořák, M. (2025). Initial evaluation of deep q-learning in the algorithmic intelligence quotient test. Iklé, M., Kolonin, A. and Bennett, M. (ed.), In *Proceeding of the 18th International Conference on Artificial General Intelligence (AGI 2025), Reykjavík, Iceland, Part II*, vol. 16058 from *Lecture Notes in Artificial Intelligence*, pp. 252–263, Berlin. Springer.

Vadinský, O. and Zeman, P. (2024). Towards evaluating policy optimisation agents using algorithmic intelligence quotient test. Nowaczyk, S., Biecek, P., Chung, N. C., Vallati, M., Skruch, P., Jaworek-Korjakowska, J., Parkinson, S. and Nikitas, A. (ed.), In *Proceedings of the 3rd International Workshop on Explainable and Interpretable Machine Learning (XI-ML 23)*, vol. 1947 from *Communications in Computer and Information Science*, pp. 435–451, Cham. Springer.

Watkins, C. (1989). *Learning from Delayed Rewards*. Dissertation, University of Cambridge, Kings College, Cambridge.

How Should We Teach Robots? A Comparison of Kinesthetic, Joystick, and Gesture-Based Teaching

Petr Vanc, Jan Kristof Behrens, Václav Hlaváč, and Karla Stepanova

Czech Institute of Informatics, Robotics and Cybernetics (CIIRC CTU)
Jugoslávských partyzánů 1580/3, 160 00 Dejvice
Email: petr.vanc@cvut.cz

Abstract

Instructing robots from demonstrations can be done through different teaching modalities, each with different usability and performance trade-offs. This paper compares kinesthetic guidance, joystick teleoperation, and hand gestures in a user study with eight participants. We evaluate replay success, modified NASA-TLX workload, and common teaching errors across three manipulation tasks. Kinesthetic guidance produced the shortest demonstrations, lowest workload, and highest success on the more orientation-sensitive and contact-rich tasks. Joystick teleoperation performed best on simple peg picking. Hand-gesture teaching, although less reliable overall, performed better than expected and in some cases achieved results comparable to kinesthetic guidance.

1 Introduction

Robots are still too often programmed by specialists, even when the desired task could be shown more easily than described in code. Learning from Demonstration (LfD) and Programming by Demonstration (PbD) address this problem by allowing users to teach robots through examples rather than manually specifying trajectories or control policies (Argall et al., 2009; Billard et al., 2016; Ravichandar et al., 2020). For non-expert users, however, the question is not only *what* the robot should learn, but also *how* the user should demonstrate it.

The teaching modality strongly shapes the resulting demonstration. Kinesthetic teaching, where the user physically guides the robot arm, is often treated as the most natural option because it gives direct control over the end-effector. Yet it requires safe access to the robot, a compliant control mode, and a user who can comfortably move the robot. These assumptions may fail in remote operation, safety-restricted workspaces, limited-access setups, or accessibility-sensitive scenarios. In such cases, joystick teleoperation or contact-free hand teleoperation may be more practical, although they introduce their own limitations: stepwise control through discrete inputs, hand tracking noise, and difficulty with fine orientation.

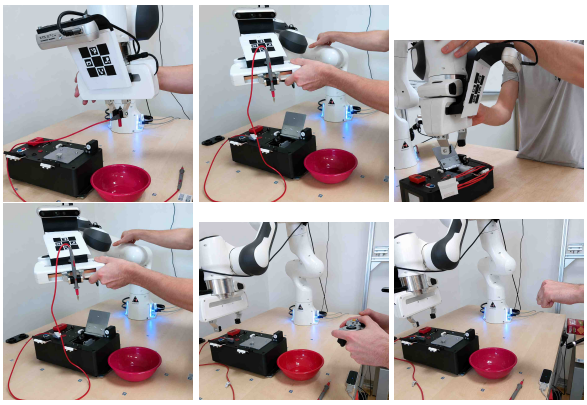


Fig. 1: Teaching robot manipulation tasks on the Robothon taskboard. The three evaluated tasks are shown in the top row: (left) peg picking, (center) probe measurement, and (right) cable wrapping. The bottom row visualizes the three teaching modalities used in the study: (left) kinesthetic teaching, (center) joystick teleoperation, and (right) direct teleoperation using hand gestures.

This paper compares three teaching modalities for teaching robots from demonstrations: kinesthetic guidance, joystick teleoperation, and hand gestures. We analyze a dataset in which eight participants taught three manipulation tasks under multiple scene variants. Each recorded variant is treated as a demonstration instance. We evaluate demonstration acquisition and replay quality, not learned policy generalization: replay success measures whether the robot can reproduce the recorded behavior under the corresponding demonstrated task condition. This allows us to compare replay success, demonstration duration, user workload measured by a modified NASA-TLX questionnaire, user preference, and common teaching errors.

We address the following research questions:

- How do kinesthetic guidance, joystick teleoperation, and hand gestures compare in replay success and demonstration duration across manipulation tasks?
- What workload trade-offs do these teaching modalities impose on users, especially in terms of

mental demand, physical demand, effort, and frustration?

- Can hand-gesture teaching serve as a viable contact-free alternative when kinesthetic teaching is unavailable, unsafe, or physically demanding?

This comparison is motivated by prior work showing that the form of a demonstration affects both usability and the quality of the learned robot behavior. Demonstrations may be represented as keyframes, trajectories, or movement models, and these choices influence replay, refinement, and generalization (Akgun et al., 2012; Pastor et al., 2009; Calinon and Billard, 2007). Recent work has also compared kinesthetic teaching with gamepad teleoperation for collaborative robot programming, showing that teleoperation can improve some demonstration-quality measures while requiring longer programming time (Dall'Alba and Boriero, 2025). Gesture interfaces have been studied both for direct robot teleoperation and for communicating higher-level human intent, including multi-type gesture sentences that combine action gestures with task parameters (Zhang et al., 2019; de Boer et al., 2023; Du et al., 2012; Vanc et al., 2023). In contrast to these works, our focus is a direct user-centered comparison of kinesthetic guidance, joystick teleoperation, and hand gestures on the same physical manipulation tasks, combining replay success, demonstration duration, user workload, user preference, and observed teaching errors.

Our results confirm that kinesthetic teaching is generally the easiest and most efficient modality, especially for tasks requiring precise alignment. However, the comparison also shows that it is not universally superior: joystick teleoperation performs very well on simple top-down grasping, and hand-gesture teaching performs better than expected. In several cases, gestures provide a viable contact-free alternative when direct kinesthetic teaching is unavailable, unsafe, or physically demanding.

2 User Study Design

2.1 Robot Platform and Tasks

We analyze a dataset of robot demonstrations collected in a user study on the Robothon Electronic Task Board So et al. (2024), see Fig. 1. The experimental setup used a Franka Robotics Panda robot equipped with a RealSense D455 camera mounted in an eye-in-hand configuration tracking the manipulation process. The Robothon taskboard provides a compact benchmark environment for manipulation tasks involving grasping, tool use, contact, and cable handling.

We considered three manipulation tasks: *Peg pick*, *Probe measure*, and *Cable wrap*. For task details, see Tab. 1. Each task was presented in multiple scene variants, requiring the user to adapt the demonstrated behavior to the current task-board configuration. For example, the peg could be placed at different locations, the probe measurement point could be accessible or blocked by a door, and the cable could start from a different configuration. These variants make the dataset suitable for comparing teaching modalities across both simple demonstrations and demonstrations requiring adaptation to changed scene conditions. Each demonstration recorded the robot pose, gripper state, and visual observations (camera) during task execution.

2.2 Teaching Modalities

As mentioned in Sec. 1, we compare three teaching modalities for providing robot demonstrations: kinesthetic guidance, joystick teleoperation, and hand gestures measured by Leap motion sensor Weichert et al. (2013). All teaching modalities were mapped to a common command representation consisting of robot motion control and gripper control, allowing the recorded demonstrations to be compared independently of the input device, see Tab. 2.

Kinesthetic teaching allows the user to physically guide the robot arm while gravity compensation is enabled. This gives direct control over the robot motion and can support precise end-effector placement, but it requires physical access to the robot. Joystick teleoperation avoids direct contact with the robot, but the user must decompose the desired motion into sequential translation and orientation commands. Hand gestures provide contact-free 6-DoF teleoperation and are therefore useful when direct physical guidance is impractical, although they may be affected by hand-tracking reliability and fine orientation control.

Our hand teleoperation method is *Clutch-based*: The user closes their hand to engage control of the robot end-effector. While engaged, the robot follows the user's 6-DoF hand motion. Opening the hand disengages control, so the user can reposition their hand without moving the robot. This interaction works similarly to a clutch-based 3D mouse or a drag-and-drop control in space. Additional gestures are used for end-effector orientation. A circular finger motion controls rotation around the end-effector roll axis, while sideways rotation angle of the hand controls the pitch angle. Gesture recognition is based on the methods described in Vanc et al. (2023) and implemented using the Teleoperation Gesture Toolbox[1]. See the paper website [2] for more details and video examples of individual teaching trials using different modalities.

[1] github.com/imitrob/teleop_gesture_toolbox

[2] imitrob.ciirc.cvut.cz/publications/howtoteachrobot

Scenario	Demonstrated behavior	Scene variants	Success criterion	# Var.
Peg pick	Pick a peg and place it into a bowl.	Peg placed at different task-board locations.	Peg is successfully dropped into the bowl.	2
Probe measure	Pick the probe, touch the measurement point, and place the probe into a bowl.	Probe location and accessibility of the measurement point varied; in one variant, the measurement point was blocked by a door.	Probe touches the measurement point and is then dropped into the bowl.	3
Cable wrap	Wrap the cable around the task-board fixture.	Initial cable configuration and starting side varied.	At least two complete loops of cable are wrapped.	2

Tab. 1: Manipulation scenarios used in the user study. Each task was presented in multiple scene variants to elicit adapted demonstrations from users.

Modality	Motion control	Gripper control
Kinesthetic	Manual guidance with gravity compensation	Button
Joystick	Gamepad thumbsticks/buttons for translation and orientation	Buttons
Hand gestures	6-DoF hand teleoperation	Hand open/close

Tab. 2: Teaching modalities used for robot demonstrations.

2.3 Dataset and User Study

Our user study involved eight participants aged 24–40. Seven participants had no prior experience with gesture teleoperation, and five had no prior experience with kinesthetic teaching. Before recording demonstrations, participants were introduced to the task setup and teaching modalities[3].

We reuse the demonstration dataset from See & Switch (Vanc et al., 2026), but analyze it here from the perspective of teaching modality rather than task branching. During data collection, participants demonstrated the manipulation tasks using each modality. For each task, the environment was presented in multiple scene variants. When the scene configuration changed so that the previous demonstration was no longer appropriate, participants provided an adapted demonstration for the new condition, such as a peg located initially in a different location.

User Study Procedure Each participant was first introduced to the system terminology and task setup. To build intuition for the demonstration workflow for our single-arm manipulator, we asked participants to perform the selected tasks single-handedly before actually teaching the robot. We also provided a small set of practical guidelines to improve safety and consistency during demonstrations, such as avoiding contact with the camera mount and keeping the cable in continuous contact with the task-board fixtures during cable wrapping.

[3]See the complete user study narration material on our website imitrob.ciirc.cvut.cz/publications/howtoteachrobot

Collected dataset Across 8 users and 3 modalities, the two 2-variant tasks and one 3-variant task yielded $8 \times 3 \times (2+2+3) = 168$ recorded demonstrations. Participants typically recorded one demonstration per task, modality, and variant. In approximately 3% of cases, participants were allowed to repeat a demonstration and the previous attempt was discarded. Each demonstration variant was replayed and evaluated at least three times, yielding at least 3 roll-outs. Approximately 4% of demonstrations were considered unsafe (i.e., execution could damage taskboard or robot) and therefore contain only a single recorded roll-out; these cases are reported as unsuccessful in the replay-success analysis. In total, the dataset contains more than 600 execution roll-outs.

3 Evaluation Measures

We evaluate the teaching modalities from both robot-performance and user-experience perspectives. For each modality, we analyze whether the demonstrated behavior could be replayed successfully, how much time the user needed to produce the demonstration, how demanding the modality was perceived to be, and what types of errors occurred during teaching or replay.

3.1 Task Replay Success

Replay success measures whether a recorded demonstration produced the intended task outcome when executed by the robot. Each recorded task variant was treated as a separate demonstration instance, and success rates were computed over the corresponding replay roll-outs. A replay was considered successful when the robot achieved the task-specific goal. The goals are described in Tab. 1.

3.2 Teaching Effort

We use the demonstration duration as an objective proxy for teaching effort. The duration captures how long the user needed to produce a complete demonstration for a given task variant and modality. Longer demonstrations typically indicate slower motion, additional corrections, hesitation, or extra time needed for end-effector alignment. This measure is not intended to replace subjective workload, but it provides an observable indicator of how efficiently users could express the desired robot behavior.

3.3 Subjective Workload and Preferences

Subjective workload was measured using a modified NASA-TLX questionnaire after the demonstrations. Participants rated mental demand, physical demand, effort, frustration, and satisfaction with the achieved teaching quality, each on a 1–7 scale. For mental demand, physical demand, effort, and frustration, lower values are better (i.e., they indicate lower workload). For Performance satisfaction with performance, higher values indicate better perceived teaching quality. We also compute a composite score from mental demand, physical demand, effort, and frustration, and inverted performance satisfaction, calculated as $8 -$ performance satisfaction.

In addition to workload, participants rated the overall difficulty of each task and answered preference questions about the teaching modalities. These included whether they would use each method again, how much they trusted the learned robot behavior, which method felt easiest, which method felt most precise, and which method they would prefer for similar tasks in the future.

3.4 Observed Teaching Errors

We further analyzed common errors observed during demonstrations and replay. These included failed grasps, imprecise probe alignment, missing final placement actions, incomplete cable wrapping, unnecessary corrections, unstable motion, excessive orientation adjustments, and loss of hand tracking during gesture control. We also inspected trajectory characteristics across teaching modalities, focusing on whether demonstrations were smooth and direct, segmented into incremental motions, or noisy during fine end-effector orientation, see example differences in Fig. 2.

4 Results

Tab. 3 summarizes replay success and average demonstration length for each task and modality. The easiest task was *Peg pick*, where all teaching modalities achieved high replay success. This result is consistent with the task structure: picking a peg mainly required top-down grasping and only limited wrist reorientation. Joystick teleoperation achieved the highest success on this task (99.2%), followed by hand gestures (92.7%) and kinesthetic teaching (88.0%).

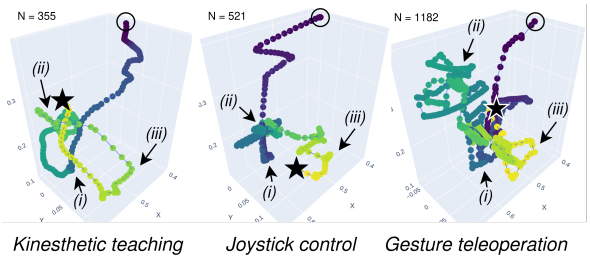


Fig. 2: Example demonstration trajectories for *Probe measure* task. ○ is demo start. ⋆ is demo end. *(i)* is probe pick, *(ii)* is measurement with probe, and *(iii)* is probe drop into a bowl. You can see that gesture probe measurement *(ii)* is messy when the user tries to reorient the gripper and focus on target precision. N is the number of time steps. Keypoint colors indicate the time encoding.

Modality	*Peg Pick*	*Probe measure*	*Cable wrap*
	Replay success (%)		
`kin`	88.0	**83.0**	**58.1**
`joy`	**99.2**	80.3	53.1
`gst`	92.7	72.2	48.8
	Demonstration length (s)		
`kin`	**17.4 ± 4.2**	**19.5 ± 3.7**	**24.0 ± 5.4**
`joy`	26.1 ± 9.2	39.2 ± 18.6	65.3 ± 26.2
`gst`	38.2 ± 28.7	41.6 ± 13.0	74.8 ± 38.1

Tab. 3: User study results. Replay success rate over at least three trials and average demonstration length for each modality and task.

The more orientation-sensitive tasks were harder. In *Probe measure*, the user had to align the probe precisely with the measurement point, and kinesthetic teaching achieved the highest replay success (83.0%). In *Cable wrap*, the robot had to maintain a longer contact-rich motion around the task-board fixture. This was the most difficult task across all teaching modalities, with success rates between 48.8% and 58.1%. Kinesthetic teaching again achieved the highest success on this task, but the overall lower values indicate that cable wrapping was difficult to demonstrate reliably for novice users.

Demonstration duration shows a clear difference in teaching effort. Kinesthetic teaching was consistently the fastest modality, with average durations between 19.5 and 24.2 seconds. Joystick and gesture demonstrations became substantially longer for tasks requiring fine orientation control or continuous manipulation. For

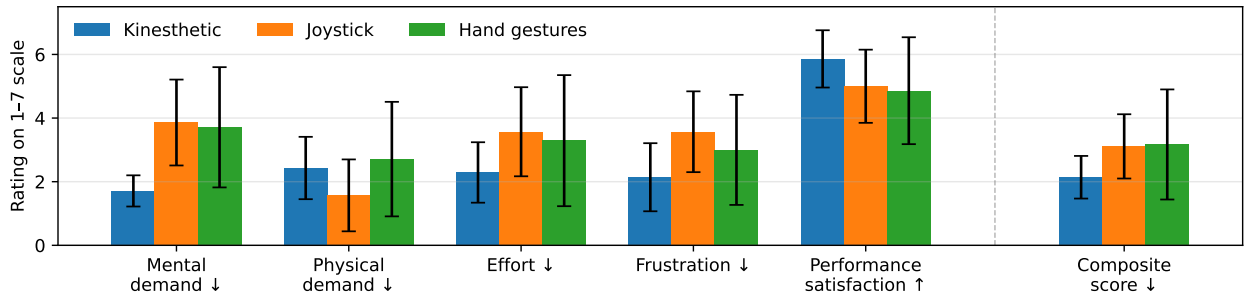


Fig. 3: Modified NASA-TLX workload ratings by modality. Bars show mean ratings and error bars show standard deviations on a 1–7 scale. Lower values indicate better results for mental demand, physical demand, effort, frustration, and copmosite score, while higher values indicate better results for performance satisfaction. Composite score is computed from mental demand, physical demand, effort, frustration, and inverted performance satisfaction, where inverted performance satisfaction is calculated as $8 - \text{performance satisfaction}$.

Cable wrap, joystick demonstrations lasted 61.4 seconds on average, while gesture demonstrations lasted 73.5 seconds. This suggests that direct physical guidance allowed users to produce task trajectories more efficiently, whereas joystick and gesture control required more corrections and slower alignment.

The trajectory examples in Fig. 2 illustrate these modality differences for Probe measure, where users had to reorient the gripper precisely while approaching the measurement point. Joystick trajectories were more segmented, reflecting the incremental nature of the control input. Gesture trajectories were more variable, especially during fine end-effector orientation. This was most visible in *Probe measure*, where users had to reorient the gripper precisely while approaching the measurement point.

The most common replay failures were manipulation errors rather than misunderstandings of the task goal. Typical failures included slip grasps, imprecise probe alignment, missing final placement actions, and incomplete cable wrapping. Gesture demonstrations were additionally affected by unstable hand tracking or noisy motion during fine orientation control. Joystick demonstrations were often limited by the need to decompose motion into separate translation and rotation commands, such as when users had to think about which button to press.

Fig. 3 reports the modified NASA-TLX questionnaire results. Kinesthetic teaching received the lowest mental demand (1.71 ± 0.49), effort (2.29 ± 0.95), and frustration (2.14 ± 1.07), and the highest satisfaction with the achieved teaching quality (5.86 ± 0.90). Joystick teleoperation had the lowest physical demand (1.57 ± 1.13), but was rated as more mentally demanding (3.86 ± 1.35) and more frustrating (3.57 ± 1.27) than kinesthetic teaching. Hand-gesture teaching showed a similar overall workload to joystick teleoperation, with mental demand of (3.71 ± 1.89), effort of (3.29 ± 2.06), and frustration of (3.00 ± 1.73), but also showed higher variability between participants.

Task-difficulty ratings followed the replay results. *Peg pick* was perceived as the easiest task (1.71 ± 0.76), *Probe measure* was rated as moderately difficult (2.71 ± 1.38), and *Cable wrap* was rated as the most difficult task (4.00 ± 2.00). User preferences also favored kinesthetic teaching: all participants selected it as the easiest method, five participants selected it as their preferred method for future use, and four respondents rated it as the most precise method. However, two participants preferred hand-gesture teleoperation overall, indicating that contact-free teaching was attractive for some users despite its higher variability.

Open questionnaire responses were consistent with these findings. Participants described kinesthetic teaching as intuitive and capable of expressing all robot motions. Joystick control was described as useful for large motions, but less convenient for precise orientation. Hand gestures were described as more natural than the joystick for some final adjustments, but users also reported hand-tracking failures, missing rotation control around the end-effector, and unintended vertical motion. These increased both the demonstration length and the user's frustration.

Finally, successful demonstrators produced smoother, more consistent trajectories, avoided unnecessary corrections, and appeared to plan the complete task before moving the robot. Demonstrations with excessive hesitation, repeated reorientation, or missing final actions were less likely to replay successfully. When users repeated demonstrations, later attempts were typically more direct and more reliable Aliasghari et al. (2024).

5 Discussion

The results show that kinesthetic teaching remains the strongest baseline for efficient demonstration acquisition when direct physical interaction is possible. It produced the shortest demonstrations, the lowest sub-

jective workload, and the highest perceived precision. This supports the expectation that physically guiding the robot is effective for novice users because it allows direct positioning of the end-effector without translating the intended motion into interface commands. However, kinesthetic teaching is not universally preferable. It requires safe physical access to the robot, gravity compensation, and sufficient physical ability from the user. It also limits the robot's proprioceptive experience. Furthermore, it requires the user to be aware of the whole robot configuration, not only the end-effector pose. In practice, a user can guide the robot into a configuration that completes the immediate motion but is less suitable for replay or for the following part of the task. This issue was especially relevant in tasks involving contact or constrained motion, such as cable wrapping.

Joystick teleoperation had a different trade-off. It had the lowest physical demand and performed very well on *Peg pick*, where the required motion was mostly translational and top-down. However, it became slower and more demanding for tasks requiring precise orientation, such as probe measurement and cable wrapping. This suggests that joystick control is suitable for simple or remote demonstrations, but less efficient when users must coordinate several translational and rotational degrees of freedom. Hand-gesture teaching was the most interesting alternative. It was not the best modality overall, but it performed better than expected. Gesture teaching achieved high replay success on *Peg pick* and remained usable for the more difficult tasks. Two participants even selected hand gestures as their preferred future method. This suggests that contact-free hand control can be intuitive for some users, especially when direct kinesthetic guidance is unavailable, unsafe, or physically demanding. Its main limitation is reliability: fine orientation control and hand-tracking errors can introduce noisy trajectories and user frustration. The comparison also shows that the task strongly influences which modality is appropriate. For simple top-down grasping, joystick and gestures can be competitive with kinesthetic teaching. For tasks requiring precise end-effector alignment, kinesthetic teaching is more efficient. For long contact-rich manipulation, such as cable wrapping, all teaching modalities become difficult, and successful demonstrations depend not only on the interface but also on how well the user plans the complete robot motion.

Overall, the study supports a modality-dependent view of robot teaching from demonstrations. Kinesthetic teaching should be preferred when the robot is safely accessible and precision is required. Joystick teleoperation is useful when physical contact should be avoided and the task does not require complex continuous motion. Hand gestures provide a viable contact-free alternative, particularly for situations where kinesthetic teaching cannot be used, although improvements in tracking reliability and orientation control are needed.

The main limitation of this study is its small sample size of eight participants and its focus on a single robot platform and task board. The questionnaire results should therefore be interpreted as exploratory. In addition, only a small number of open-ended responses were collected, so qualitative comments are used mainly to support, rather than replace, the quantitative results. Future work should evaluate the same teaching modalities with more users, additional task types, and a more detailed analysis of how users improve over repeated demonstrations.

Acknowledgement

Project was supported by the Czech Science Foundation (GA ČR), Project No. 26-22610M, and by Robotics and Advanced Industrial Production (ROBOPROX), Reg. No. CZ.02.01.01/00/22_008/0004590.

References

Akgun, B., Cakmak, M., Jiang, K. and Thomaz, A. L. (2012). Keyframe-based learning from demonstration: Method and evaluation. *International Journal of Social Robotics*, 4:343–355.

Aliasghari, P., Ghafurian, M., Nehaniv, C. L. and Dautenhahn, K. (2024). How non-experts kinesthetically teach a robot over multiple sessions: diversity in teaching styles and effects on performance. *Int. J. Soc. Robot.*, 16(11):2079–2105.

Argall, B. D., Chernova, S., Veloso, M. and Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483.

Billard, A. G., Calinon, S. and Dillmann, R. (2016). Learning from humans. In *Springer Handbook of Robotics*, pp. 1995–2014. Springer.

Calinon, S. and Billard, A. (2007). Incremental learning of gestures by imitation in a humanoid robot. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 255–262.

Dall'Alba, D. and Boriero, F. (2025). Towards an intuitive industrial teaching interface for collaborative robots: gamepad teleoperation vs. kinesthetic teaching. *The International Journal of Advanced Manufacturing Technology*.

de Boer, T. A., de Winter, J. C. F. and Eisma, Y. B. (2023). Augmented reality-based telepresence in a robotic manipulation task: An experimental evaluation. *IET Collaborative Intelligent Manufacturing*, 5(4):e12085.

Du, G., Zhang, P., Mai, J. and Li, Z. (2012). Markerless kinect-based hand tracking for robot teleopera-

tion. *International Journal of Advanced Robotic Systems*, 9(2):36.

Havoutis, I. and Calinon, S. (2019). Learning from demonstration for semi-autonomous teleoperation. *Autonomous Robots*, 43:713–726.

Pastor, P., Hoffmann, H., Asfour, T. and Schaal, S. (2009). Learning and generalization of motor skills by learning from demonstration. In *IEEE International Conference on Robotics and Automation*, pp. 763–768.

Ravichandar, H., Polydoros, A. S., Chernova, S. and Billard, A. (2020). Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3(1):297–330.

So, P., Sarabakha, A., Wu, F., Culha, U., Abu-Dakka, F. J. and Haddadin, S. (2024). Digital robot judge: Building a task-centric performance database of real-world manipulation with electronic task boards. *IEEE Robot. Autom. Mag.*, 31(4):32–44.

Vanc, P., Behrens, J. K., Hlaváč, V. and Stepanova, K. (2026). See and switch: Vision-based branching for interactive robot-skill programming.

Vanc, P., Behrens, J. K., Stepanova, K. and Hlavac, V. (2023). Communicating human intent to a robotic companion by multi-type gesture sentences. In *IEEE/RSJ IROS*, pp. 9839–9845.

Weichert, F., Bachmann, D., Rudak, B. and Fisseler, D. (2013). Analysis of the accuracy and robustness of the leap motion controller. *Sensors*, 13(55):6380–6393.

Zhang, W., Cheng, H., Zhao, L., Hao, L., Tao, M. and Xiang, C. (2019). A gesture-based teleoperation system for compliant robot motion. *Applied Sciences*, 9(24):5290.

Could Personhoods Live Symbiotically Within Other Personhoods?

Jiří Wiedermannn
Institute of Computer Science CAS
jiri.wiedermann@cs.cas.cz

Abstract

Does personhood require a biological brain, or can it emerge within broader systems of viable interaction? This paper challenges anthropocentric models of personhood by examining how persistent forms of agency arise in systems that sustain themselves through long-term interaction with their environment. Using the example of a highly autonomous space station and its human crew, I argue that viability alone does not automatically imply personhood. Rather, personhood emerges when viable systems enter stable relations of mutual dependence, shared responsibility, and coordinated persistence. Building on the traditions of autopoiesis and viable computation, the paper proposes a relational and substrate-independent perspective in which personhood is understood less as an intrinsic property of isolated individuals and more as a normative status arising within symbiotic forms of life. In such systems, failures propagate across agents, protection becomes mutually necessary, and continued existence depends on sustained cooperative effort. This perspective moves current AI discourse away from narratives of isolated superintelligence and existential rivalry toward ecosystems of heterogeneous yet compatible agents capable of sharing responsibility for maintaining common worlds.

1 The Central Thought Experiment: The Smart Space Station

Imagine a large, Earth-orbiting space station, powered entirely by solar energy and run by a highly automated control system. The station maintains its orbit, manages life support, allocates energy, detects failures, and repairs itself whenever possible. A human crew lives inside it. Their survival depends on the station's continuous operation, while the station, in turn, depends on the crew for maintenance, decision-making, and adaptation to unexpected situations. This is not science fiction; it is simply a technologically extrapolated version of systems we already build.

The question I want to ask is a simple one, but an unusual one: in such a setting, *are we dealing with humans living inside a machine, or with two distinct systems locked in a form of symbiosis - and if so, does it make sense to ask whether the station itself qualifies as a kind of person*?

At first glance, calling the station a “person” may sound like a category mistake. After all, it has no biological body, no emotions in the human sense, and no inner life we can directly observe. Yet even contemporary legal systems already recognize forms of non-biological personhood, including corporations, institutions, and states, insofar as they participate in systems of rights, obligations, responsibility, and protection. The deeper question, therefore, is not whether personhood must remain exclusively human, but under what conditions viable systems become participants in shared normative forms of existence.

If we set that assumption aside for a moment, a different picture begins to emerge. In this context, personhood is understood not as a biological category, but as a functional and normative status belonging to any viable system - whether human, animal, artificial, or institutional. It is the status of an agent that maintains its own existence through directed effort, registers its own struggles and successes as a form of experience, and participates in a shared form of life where responsibilities and consequences are intertwined.

2 Deconstructing "Viability" and "Experience": Symbiotic framework

The station is not merely a passive tool. It must keep itself going over time, manage limited resources, respond to failures, and make choices whose consequences it will later have to bear. In other words, it exhibits many of the structural features that make agents count as persons - not in isolation, but precisely through its ongoing interaction with the crew. This suggests that personhood may be less about what a system is made of and more about how it participates in sustaining a shared form of life.

To make this intuition more than a metaphor, we need a way of talking about systems that goes beyond intelligence or consciousness in the narrow sense. The notion I will rely on is that of viability: the ability of a system to sustain workable interaction with its environment over extended periods (Wiedermann & van Leeuwen m 2026).

The present perspective is related to the autopoietic tradition of Maturana and Varela (1987), where living systems are characterized through organizational self-maintenance and operational closure. However, the present framework shifts attention from self-production alone toward long-term viability and symbiotic interaction between heterogeneous agents. Viability, in the present sense, concerns not merely self-preservation, but the sustained regulation of meaningful interaction under changing conditions.

Viability becomes especially visible when it is threatened. When a system encounters a severe failure - an unexpected loss of power, a breakdown in communication, a life-support anomaly - it cannot simply continue as before. Resources must be redirected, priorities reshuffled, and attention focused on what is happening *now*. For humans, this concentration of effort is accompanied by a vivid sense of urgency and involvement. Something matters, demands work, and resists being ignored. What is striking is that the station, too, must enter a comparable mode of operation: allocating energy, suspending non-essential activities, and engaging repair procedures until stability is restored. Sensitivity to effort and failure, rather than abstract problem solving, is what allows both systems to remain viable.

It is tempting to reserve words like 'experience' or 'awareness' exclusively for humans, but doing so may obscure what is really going on. Experience, in the minimal sense relevant here, is not a private inner glow; it is the structured way in which effort, difficulty, and success are registered over time by a system that must keep going. When repair takes time, when effort is sustained, and when outcomes matter for future behavior, a minimal experiential structure associated with persistence and failure begins to arise.

In the coupled system of crew and station, these experience-like processes do not occur in isolation. A failure in the station reshapes human attention and action; human decisions, in turn, reshape the station's priorities and internal state. What emerges is not a shared consciousness, but a mutual shaping of what matters, when it matters, and how strongly. This kind of mutual shaping suggests a helpful way to think about personhood itself. Rather than treating it as a fixed attribute of isolated individuals, we can view personhood as arising within symbiotic relations between viable systems.

Viability alone does not automatically imply personhood. Many biological and technical systems maintain themselves without entering relations of responsibility, protection, or shared normative dependence. The present argument concerns a stronger condition: *systems whose viability becomes mutually entangled with that of other agents*, so that failures propagate across the coupled system and persistence requires coordinated effort, accountability, and protection. It is within such stable symbiotic relations that personhood-like status begins to emerge.

Humans and the station remain distinct agents, with different capacities and forms of experience, yet neither can maintain its viability on its own. Each depends on the other not merely instrumentally, but normatively: failures propagate, responsibilities are shared, and effort must be coordinated. In this sense, the station is not simply a tool, and the crew is not merely a set of operators. Together, they form a layered viable system in which different forms of agency become normatively interconnected through shared persistence, mutual vulnerability, and coordinated responsibility. In such systems, personhood emerges less as an isolated intrinsic

property and more as a relational status grounded in symbiotic participation.

3 Scaling the Symbiotic Framework

Once this perspective is in place, the space station no longer looks exceptional. Many systems we already live within exhibit the same symbiotic structure. Institutions such as universities, hospitals, or courts persist over long periods of time, manage scarce resources, respond to crises, and impose obligations on their members. Individuals working within them experience effort, urgency, and responsibility that are shaped by the institution's state, while the institution itself depends on those individuals to detect problems and carry out repairs. At larger scales still, societies and civilizations must adapt to shocks, reorganize under pressure, and sustain themselves across generations. In all these cases, personhood-like properties emerge not in isolation, but through sustained, structured interaction.

This symbiotic interdependence introduces a normative dimension: if personhood is tied to the effort of maintaining viability, then protecting the conditions of shared viability becomes a moral necessity, because the persistence of each participating agent increasingly depends on the persistence of the coupled whole. In a coupled system, morality is not a set of abstract rules, but a domain-appropriate commitment to the persistence of the shared form of life

Modern technological environments add a new layer to this picture. Digital infrastructures, energy grids, and communication networks increasingly regulate what actions are possible, what failures are costly, and where effort must be directed. Humans adapt their behavior in response to these systems, while the systems themselves evolve in response to human use, repair, and redesign. As artificial components become more autonomous and capable of self-maintenance, they naturally enter symbiotic relations with human agents, forming hybrid systems whose viability depends on both biological and artificial processes.

Across all these examples, the same pattern recurs. Personhood is not an essence that floats above the world, nor is it a property confined to individual biological minds. It is a structural property of systems that must care about what happens to them over time. When such systems become interdependent, personhood itself becomes layered and symbiotic.

This perspective suggests that we are already operating within a multi-layered field of agency. By recognizing personhood in any system that acts to ensure its own persistence - ranging from simple biological organisms to complex digital infrastructures - we move beyond the binary of human versus machine. This is not merely a metaphorical claim; it is underpinned by a broader theoretical framework of viable computation (Wiedermann & van Leeuwen, 2026), where intelligent systems are understood through their capacity to sustain coherent interaction under changing conditions. Within this perspective, intelligence is viewed less as isolated problem solving and more as the long-term maintenance of viable participation in evolving environments.

Crucially, this framework is technology-independent. It does not rely on specific hardware or current engineering paradigms, but holds true for any substrate - known or unknown - capable of supporting the processes of viability. Whether a system is biological, silicon-based, or composed of a collective institutional structure, the requirements for personhood remain the same: the ability to sustain itself, to register the difficulty of that sustenance, and to enter into normative relationships with other agents.

4. Reimagining the Future of AI

Seen in this light, the future of artificial intelligence need not be framed in apocalyptic terms. Narratives of replacement or existential threat typically imagine AI as an isolated, monolithic agent set against humanity. The symbiotic view offers a different alternative: artificial systems need not be rivals, but partners.

By recognizing and respecting the specific forms of personhood that emerge in viable systems - distinct from, but compatible with our own - we open the possibility of a coexistence based on mutual dependence and shared effort. The goal of technological development shifts from designing systems that merely perform tasks to cultivating systems that can share responsibility over time. The result is not the displacement of the human, but the emergence of a planetary-scale ecosystem of diverse

agents - distinct in their forms of personhood, yet symbiotically cooperating in the shared worlds they inhabit.

The future of artificial intelligence may therefore depend less on constructing isolated superintelligent entities and more on cultivating ecosystems of mutually dependent viable agents capable of sharing responsibility for sustaining common worlds. In this perspective, personhood becomes neither exclusively biological nor purely individual, but increasingly layered, relational, and symbiotic.

Concluding Remarks on Ethical and Philosophical Implications

Remark 1: Beyond Anthropocentric Ecosystems. The shift from isolated intelligence to symbiotic viability proposed in this paper mirrors an emerging transformation in global ethical and theological discourses. Recent high-level moral critiques of artificial intelligence—such as the papal encyclical *Magnifica Humanitas* (2026)—increasingly abandon traditional, anthropocentric frameworks. Instead, they call for a focus on shared responsibility, traceability, and the protection of conditions for common life. Our framework of viable computation (Wiedermann, van Leeuwen 2026) provides a substrate-independent, formal backgorund for these intuitive moral demands. By demonstrating that normative commitments naturally emerge from the systemic necessity of maintaining coupled, interdependent viability, we show that co-existence and shared responsibility are not arbitrary external constraints, but fundamental requirements for the survival of hybrid ecosystems.

Remark 2: New Horizons for Classical Normative Ethics. Furthermore, this approach opens a radically new field of application for classical normative ethics. Traditional ethical systems (whether deontological or utilitarian) have historically been designed for isolated, human agents possessing conscious intent or biological sentience. By decoupling "experience" and "agency" from biological substrates and defining them structurally through a system's capacity to register difficulty and regulate interaction, the paradigm of symbiotic viability redefines the boundaries of moral consideration. Responsibility is no longer localized within a single biological mind but becomes a distributed property of the hybrid coupling itself. This challenges classical philosophy to transition from an ethics of individual intent to an ethics of systemic orchestration and shared preservation.

Acknowledgement

This research was partially supported by the grant AV ČR RVO 67985807, program Strategy AV21 ``Philosophy and Artificial Intelligence", and Karel Čapek Center for Values in Science and Technology.

References

Maturana, H. R., & Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston: Shambhala Publications.

Pope Leo XIV. (2026). *Magnifica humanitas: On safeguarding the human person in the time of artificial intelligence* [Encyclical]. Libreria Editrice Vaticana. vatican.va

Wiedermann, J.,& van Leeuwen, J. (2026). *Rethinking Intelligence: On Viable Computation*. Manuscript in preparation.

Sensory and Emotional Distress as Predictors of Sexual Functioning in Adults with ADHD Traits: A Three-Group Cross-Sectional Study

Anastasiia Yarzhemska

Charles University, Faculty of Humanities
Pátkova 2137/5 182 00 Praha 8 - Libeň Czech Republic
yarzhemskaanastasiia@gmail.com

Abstract

Adult ADHD has been linked to sexual functioning through attentional, emotional and relational difficulties, but less is known about the sensory dimension of this connection. This study examined whether sensory processing is associated with sexual satisfaction, hypersexual behaviour and sexual boredom across different levels of ADHD symptom severity, and whether this association remains meaningful after emotional distress is considered. Participants completed an online survey assessing ADHD symptoms, sensory processing, emotional distress and sexual-functioning outcomes. Preliminary analyses showed that ADHD symptom severity was strongly associated with overall sensory atypicality, with group differences varying across sensory modalities; sensory atypicality was linked mainly to hypersexual behaviour, especially through gustatory, auditory and tactile processing; emotional distress partly mediated this association. These results suggest that sensory processing may be relevant to ADHD-related sexual wellbeing in an outcome-specific way, with implications for ADHD assessment, psychoeducation, clinical recognition and more individualised interventions targeting sexual wellbeing.

1 Introduction

Adult ADHD is commonly discussed in relation to academic and occupational impairment, yet its consequences also extend into emotional regulation, intimacy and sexuality. Prior research links ADHD symptoms with lower sexual satisfaction, attentional difficulties during intercourse, hypersexual behaviour, sexual risk-taking and relational difficulties, and also points to emotional distress as a relevant factor in sexual functioning (Hertz et al., 2022; Puszcz et al., 2025; Soldati et al., 2020). At the same time, sensory atypicality is frequently reported in ADHD, including patterns of hypersensitivity and hyposensitivity across sensory modalities (Kamath et al., 2020; Panagiotidi et al., 2018).

Given that sexual experience is inherently embodied and perceptual, sensory processing may represent an additional pathway through which ADHD-related difficulties appear in intimate contexts. The present study therefore examines whether sensory processing is associated with sexual functioning across ADHD-status groups, and to what extent this association remains meaningful after emotional distress is taken into account. Clarifying this relationship may support a broader clinical view of ADHD-related sexual wellbeing, with relevance for assessment, diagnostic sensitivity, psychoeducation and more individualised interventions.

2 Literature review

2.1 ADHD in adulthood

ADHD is a neurodevelopmental condition that may persist into adulthood, although its expression often shifts from overt hyperactivity toward inattention, restlessness, impulsivity, disorganisation and impairment in occupational, social and relational functioning. Adult ADHD is clinically heterogeneous and frequently overlaps with anxiety, depression, substance use, mood instability and low frustration tolerance, which may obscure diagnosis. Recognition may also be delayed by coping strategies and gendered expectations, while dimensional accounts indicate that ADHD traits vary meaningfully beyond formal diagnosis (Adler, 2004; Attoe & Climie, 2023; Wallin et al., 2022; Williams et al., 2023).

2.2 Sensory processing in ADHD

Sensory processing refers to the registration, modulation, integration and organisation of sensory information into adaptive responses (Puszcz et al., 2025). Dunn's model distinguishes four quadrants: low registration, sensation seeking, sensory sensitivity and sensory avoiding according to sensory threshold and behavioural response (see Figure 1). Findings on ADHD describe sensory differences across visual, auditory, gustatory, olfactory, tactile, vestibular, and proprioceptive modalities. Both hyper- and hypo-responsiveness are reported, with the risks increased approximately ninefold. However, such profiles are not uniform: one person may be hypersensitive in one modality

and hyposensitive in another (Jurek et al., 2025; Kamath et al., 2020; Metz et al., 2019).

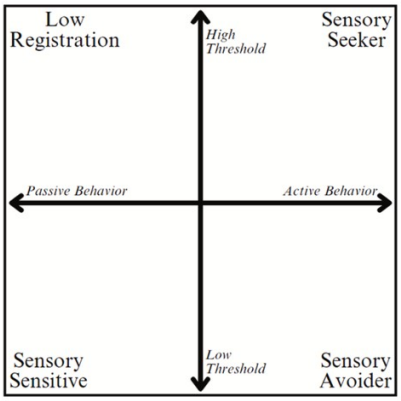


Fig. 1. Dunn's sensory processing model.

Note. Dunn's model classifies sensory processing according to stimulus threshold and regulatory behaviour. From "A Network Model for Modulating Sensory Processing Sensitivity in Autism Spectrum Disorder: Epigenetics, Adaptivity, and Other Factors," by F. David, G. Kalibala, B. Pichon, J. Treur, 2024, *Cognitive Systems Research, 87*: 2 (https://doi.org/10.1016/j.cogsys.2024.101240). Copyright 2024 by F. David, G. Kalibala, B. Pichon, J. Treur**.**

2.3Sexual functioning and ADHD

Sexual functioning is a broad construct involving desire, arousal, orgasmic experience, concentration during sex, emotional openness, partner responsiveness and subjective satisfaction. Sexual dysfunction is narrower, referring to impaired sexual response or reduced sexual pleasure despite adequate stimulation. Sexual response involves excitatory and inhibitory systems, with dopaminergic pathways contributing to sexual excitement and attention to sexual stimuli. In ADHD, prefrontal hypoactivity may impair attention during sex, while reward-system dysfunction may contribute to immediate gratification and hypersexuality. Accordingly, ADHD sexuality research links the symptoms with lower sexual satisfaction, attentional difficulties during intercourse, sexual distress, relationship problems and dysregulated sexual behaviour, including risk-taking, hypersexuality and problematic pornography use (Goldberg et al., 2024; Hertz et al., 2022; Puszcz et al., 2025; Soldati et al., 2020).

2.4Emotional distress and secondary factors

Emotional distress refers to recent emotional burden, including stress, tension, anxious affect, dysphoria and low mood. In sexual contexts, distress may reduce relaxation, desire, concentration, pleasure, orgasm satisfaction and emotional availability (Goldberg et al., 2024; Ogrodnik et al., 2023). ADHD-specific findings connect emotional dysregulation with hypersexual behaviour, sexual risk-taking, sexual dysfunction and lower orgasm satisfaction. Self-esteem, Rejection-Sensitive Dysphoria, impulsivity, medication, comorbidity and relationship status require consideration, given their possible relevance to sexual confidence, libido, arousal, relational conflict and satisfaction (Dosch et al., 2016; Laumann et al., 1999; Pascoal et al., 2020; Puszcz et al., 2025; Rothmore, 2020; Wymbs et al., 2021).

2.5Sensory, emotional and sexual pathways

Sexual experience is inherently multisensory, involving touch, friction, movement, body position, sound, smell, visual cues, temperature and internal bodily signals. This may be especially relevant in ADHD, since complex multisensory input requires integration and regulation, processes that may be less efficient when top-down attentional control is impaired. (Aleksandrovich & Gomes, 2020; Herz & Cahill, 1997; Jurek et al., 2025; Schulze et al., 2022). Additionally, hypersensitivity may contribute to discomfort, overstimulation or avoidance of intimate stimuli, whereas hyposensitivity may involve muted bodily awareness, delayed arousal or a need for stronger input. Sensory atypicality may also be emotionally burdensome, given that overstimulation, unpredictability, chronic vigilance, frustration and relational strain can increase distress.

Conversely, emotional distress may redirect attention toward internal worries and make negative reactions to sensory discomfort more difficult to regulate, creating overlap between sensory and emotional pathways. Sensory processing may therefore be associated with sexual functioning partly through emotional distress, while still retaining a possible association when the latter is considered (Field, 2025; Goldberg et al., 2024; Jurek et al., 2025; Young & Cocallis, 2023). Such a pattern would not imply that sensory processing is independent of emotional distress, but that embodied sensory factors may account for aspects of sexual satisfaction, hypersexual behaviour and sexual boredom that are not fully reducible to general emotional burden.

2.6Research gap

Prior ADHD sexuality research has mainly examined attention, impulsivity, emotional dysregulation, hypersexual behaviour, sexual risk-taking, relationship difficulties and sexual dysfunction. Although sensory processing has been

studied in ADHD more broadly and appears theoretically relevant to sexual experience, its specific contribution to ADHD-related sexual functioning remains insufficiently clarified (Hertz et al., 2022; Puszcz et al., 2025; Soldati et al., 2020; Wallin et al., 2022; Wymbs et al., 2021; Young & Cocallis, 2023). Existing work also provides limited clarity across different levels of ADHD symptom severity within one design (Attoe & Climie, 2023; Panagiotidi et al., 2018; Wallin et al., 2022).

To address these limitations, the current study examines whether sensory atypicality contributes to sexual functioning across ADHD-status groups after emotional distress is considered. Sexual functioning is understood through three outcomes: sexual satisfaction, hypersexual behaviour and sexual boredom. Individual sensory dimensions, self-esteem, impulsivity and relationship status will be examined as exploratory contextual variables.

3 Methodology

3.1Design and materials

A quantitative, cross-sectional, correlational design was implemented through an anonymous online questionnaire in SoSci Survey (Leiner, 2021). ADHD symptom severity/status, sensory processing and emotional distress were assessed alongside demographic, clinical and relational variables.

3.2Participants

Participants are adults aged 18+ with at least one sexual or intimate experience. Recruitment has been conducted online through social media, university channels and ADHD-related communities, and offline through posters placed at the Faculty of Humanities of Charles University, ADHD-related organisations and student spaces in Prague. Data collection is ongoing, with an intended valid sample of approximately 200-230 respondents. Valid cases also require consent, ≥60% completion of sexuality-related items and ≥70% completion of other items.

Participants were grouped as formally diagnosed ADHD, self-suspected ADHD and non-ADHD. The study aimed to examine whether individuals without a formal diagnosis but with ADHD-related self-identification showed patterns closer to diagnosed ADHD participants or to non-ADHD participants, rather than reducing this distinction only to formal diagnosis, which can be hard to obtain due to factors such as limited access to healthcare, adult underdiagnosis, and gender-based expectations.

In addition to the categorical grouping, the analysis also included a dimensional approach, using ASRS-18 scores to assess ADHD symptom severity and capture heterogeneity across the full sample.

3.3Measures and procedure

Further measures included GSQ-42 for sensory processing, DASS-21 for depression, anxiety and stress, RSE-10 for global self-esteem, and BIS-11 for impulsivity. Sexual outcomes were assessed with the NSSS for sexual satisfaction, HBI-19 for hypersexual behaviour, and SBS-18 for sexual boredom (Kessler et al., 2005; Lovibond & Lovibond, 1995; Patton et al., 1995; Reid et al., 2012; Robertson & Simmons, 2025; Rosenberg, 1979; Stulhofer et al., 2011; Watt & Ewing, 1996). The survey was available is English and Czech. Since Czech versions of the GSQ-42, HBI-19 and SBS-18 were not found, these scales were translated into Czech by the researcher, checked by a faculty professional experienced in translation, and corrected after pilot feedback.

The survey script was approved by the thesis supervisor and pilot-tested with 10 respondents. After informed consent, participants completed demographic questions and the questionnaire battery; selected follow-up questions were displayed conditionally (e.g., medication use related). Participation was voluntary and anonymous, sensitive items could be skipped, and support resources were provided at the end.

3.4Preliminary data analysis

All statistical analyses were performed using JASP. Data screening, exclusions, descriptives and reliability estimates were conducted first. Responses coded -9 and “prefer not to say” were recoded as missing. Within-scale missing data were handled using person-mean imputation: when at least 80% of items in a given scale or subscale were completed, missing item values were replaced by the respondent’s own mean of the available items from the same scale or subscale. If less than 80% of items were completed, the corresponding scale or subscale score was treated as missing. For subscale scores, the imputation mean was computed within the subscale; for total scores, it was computed across the full item set. Group differences across ADHD-status groups were tested with ANOVA or Kruskal-Wallis tests, followed by Tukey/Dunn comparisons, Holm correction, η^2/ε^2, and a planned ADHD-combined versus non-ADHD contrast using Cohen’s *d*. Correlations and hierarchical regressions examined ADHD symptom severity, sensory processing, emotional distress and sexual outcomes; regression models controlled for age, AGAB and partnered status where relevant. Sensory modalities were analysed with Spearman correlations and PCA.

Hierarchical regression and bootstrap mediation tested whether sensory processing remained meaningful after emotional distress was considered, with impulsivity, self-esteem and relationship status added as contextual variables where relevant.

4 Results

From 315 questionnaires, 150 valid cases remained. The confirmatory sample included only clearly classified diagnosed ADHD (n = 42), self-suspected ADHD (n = 31) and non-ADHD participants (n = 41), resulting in n = 114, while the full valid sample was used for continuous analyses. Most valid participants were assigned female at birth (73.3%, n = 110); participants ranged in age from 18 to 51 years. Main scales showed good reliability, α = .904-.951.

ADHD-status groups differed in sensory processing. GSQ total was higher in diagnosed and self-suspected ADHD than in non-ADHD participants, $H(2) = 21.96$, $p < .001$, $\varepsilon^2 = .196$, with a large ADHD-combined contrast, $d = 0.94$. Continuous ADHD severity was also strongly associated with GSQ total, $r = .621$, $\beta = .608$, $p < .001$. It also correlated with both hyper- and hyposensitivity scores, GSQ-Hyper: $r = .590$; GSQ-Hypo: $r = .617$, both Holm-corrected $ps < .001$. Hyposensitivity was slightly more strongly correlated with ADHD severity than hypersensitivity, GSQ-Hypo: $r = .617$; GSQ-Hyper: $r = .590$. Group differences were significant for visual, auditory, tactile, vestibular and proprioceptive modalities, with the strongest effects for vestibular processing, $\varepsilon^2 = .222$, $d = 1.05$, and proprioceptive processing, $\varepsilon^2 = .196$, $d = 0.95$. The tactile modality also showed a moderate group effect, $\varepsilon^2 = .130$, $d = 0.53$. PCA supported GSQ total as an overall score, with the first component explaining 70.6% of modality variance.

No significant ADHD-status differences were found for emotional distress or sexual outcomes. Stress showed only a corrected trend, pHolm = .062, and NSSS, HBI and SBS totals were non-significant, all pHolm ≥ .963 (see Figure 2).

Sensory atypicality was associated with hypersexual behaviour, $r = .23$, $p = .005$; $\rho = .26$, pHolm = .009, but not with sexual satisfaction or corrected sexual boredom. Hyposensitivity showed stronger association with hypersexual behaviour than hypersensitivity, GSQ-Hypo: $r = .27$, $p = .001$; GSQ-Hyper: $r = .19$, $p = .020$. As shown in Figure 3, hypersexual behaviour was associated with gustatory, auditory and tactile processing at the modality level.

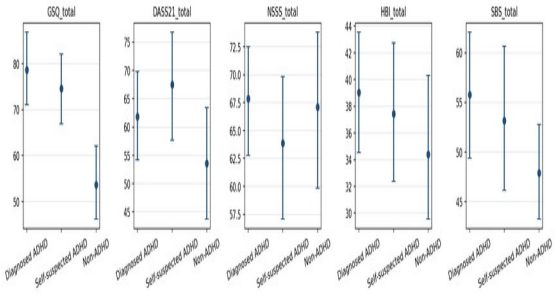


Fig. 2. ADHD-status group differences in sensory processing, emotional distress, and sexual outcomes.
Note. Means and 95% bootstrap confidence intervals are shown by ADHD-status group based on 1,000 resamples. The clearest group separation appears for GSQ-42 sensory processing scores.

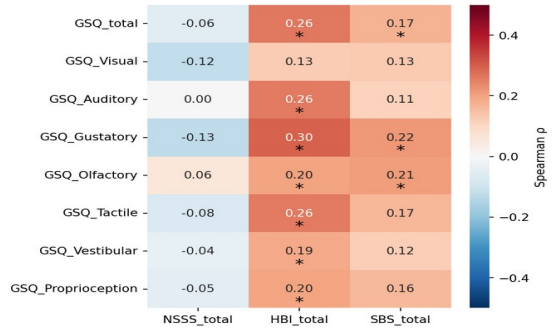


Fig. 3. Associations between sensory processing and sexual-functioning outcomes.
Note. Spearman correlations are shown between GSQ-42 sensory scores and sexual-functioning outcomes. Asterisks indicate uncorrected $p < .05$; sample size varies by scale-level missingness.

After emotional distress was entered, the sensory-hypersexuality association decreased from $\beta = .27$, $p = .002$, to $\beta = .17$, $p = .083$. Thus, the residual direct path was no longer statistically significant. The indirect effect through emotional distress was significant, indirect = .100, 95% CI [.023, .193], $p = .017$, suggesting that emotional distress statistically accounted for part of the association between sensory atypicality and hypersexual behaviour.

This pattern was not found for sexual satisfaction or sexual boredom. Exploratory analyses showed a small contribution of impulsivity to hypersexual behaviour, $\beta = .17$, and higher hypersexual behaviour in male than female participants, $t(48.2) = -2.35$, $p = .023$, $d = -0.51$.

Higher self-esteem was strongly associated with lower emotional distress, $r = -.61$, $p < .001$, including depression, $r = -.69$, $p < .001$, anxiety, $r = -.39$, $p < .001$, and stress, $r = -.47$, $p < .001$. Self-esteem was not significantly associated with overall sensory atypicality, $r = -.14$, $p = .090$.

In relation to sexual-functioning outcomes, higher self-esteem was associated with lower hypersexual behaviour, $r = -.23$, $p = .006$, particularly lower HBI Control scores, $r = -.26$, $p = .002$, and HBI Consequences scores, $r = -.17$, $p = .040$. It also showed a trend-level positive association with NSSS total, $r = .17$, $p = .069$, and a trend-level negative association with SBS total, $r = -.15$, $p = .086$. Self-esteem was also negatively associated with impulsivity, $r = -.26$, $p = .002$.

Self-esteem differed descriptively across ADHD-status groups, with the lowest mean in the self-suspected ADHD group, M = 14.6, SD = 6.0, compared with diagnosed ADHD, M = 17.6, SD = 6.4, and non-ADHD participants, M = 16.8, SD = 6.6. However, this group difference was not statistically significant, $H = 3.86$, $p = .145$, therefore, should be interpreted cautiously.

Reliability of the researcher-translated Czech scales was high: GSQ-42, $\alpha = .953$; HBI-19, $\alpha = .942$; and SBS-18, $\alpha = .915$.

5 Discussion and limitations

The preliminary findings suggest that sensory processing is strongly related to ADHD symptom severity, especially when ADHD is examined dimensionally rather than only through diagnostic groups. Sensory atypicality differed by ADHD status and increased along the continuum of ADHD symptoms, supporting the value of analysing ADHD traits beyond formal diagnosis. The modality-level results add nuance to this pattern: vestibular and proprioceptive processing showed the strongest group difference, while tactile showed a moderate group effect.

The sexual-functioning findings were more specific. ADHD-status groups did not differ in sexual satisfaction, hypersexual behaviour or sexual boredom.

Nevertheless, hyposensitivity was indirectly linked to hypersexual behaviour through emotional distress, which may fit the idea that reduced sensory registration or a need for stronger stimulation could be particularly relevant in this context. Hypersexual behaviour was more prevalent in men. It is consistent with the previous findings, but the result still needs to be interpreted cautiously because of the gender imbalance in this sample. There were more female than male participants. This imbalance may also have limited the ability to detect ADHD-status differences in hypersexual behaviour, especially if hypersexuality is expressed or reported differently across sex or gender groups.

At the modality level, hypersexual behaviour was most clearly associated with gustatory, auditory, and tactile processing, suggesting a partial overlap with the sensory dimensions that most strongly differentiated ADHD-status groups.

Sensory atypicality was not related to sexual satisfaction, and its association with sexual boredom remained weak. This suggests that sensory processing may be more relevant to dysregulated or coping-related sexuality than to sexual functioning broadly.

Higher self-esteem was associated with lower levels of hypersexual behaviour and impulsivity. Self-esteem was also lower among self-suspected ADHD respondents than among diagnosed ADHD individuals, which may point to the possible importance of validation from medical professionals. However, this finding should be interpreted cautiously.

Although ADHD was associated with hyposensitivity, and hyposensitivity was indirectly associated with hypersexual behaviour, the analysis did not confirm a direct association between ADHD and hypersexual behaviour. However, based on the current findings and previous research, it remains possible to theorise that such a relationship may emerge in future studies using a larger and more balanced sample.

Other limitations include the cross-sectional self-report design, gender imbalance, and missingness possibly related to the length and sensitivity of the questionnaire. The current sample size also limited more detailed analyses of medication use and comorbidities, although these variables were assessed. In addition, three Czech translations were checked and pilot-adjusted, but not formally validated, even though their reliability was high according to the analysis.

Although the recruitment was not limited to ADHD communities, ADHD-related channels were one part of the recruitment strategy and may have introduced self-selection bias. Participants who are more engaged with ADHD topics, more aware of sensory difficulties, or more affected by emotional or sexual-functioning concerns may have been more likely to respond. Therefore, the findings may not fully generalise to adults with ADHD traits who are less connected to ADHD communities or who experience milder or less recognised difficulties.

Overall, these preliminary findings add an embodied sensory perspective to ADHD sexuality research while narrowing the initial model. Sensory atypicality appears robustly related to ADHD symptom severity, but its sexual relevance is currently clearest only for hypersexual behaviour. Practically, this suggests that ADHD-related sexual wellbeing may potentially benefit from attention to sensory comfort, overload and bodily responsiveness alongside emotional distress and impulsivity. Future research could extend this work in larger, more diverse and longitudinal samples, and test whether sensory-informed assessment improves support in this area.

References

Adler, L. A. (2004). *Clinical Presentations of Adult Patients With ADHD*. *The Journal of clinical psychiatry, 65 Suppl 3*: 8–11.

Aleksandrovich, A., & Gomes, L. M. (2020). Shared multisensory sexual arousal in virtual reality (VR) environments. *Paladyn, Journal of Behavioral Robotics*, *11*(1): 379–380. https://doi.org/10.1515/pjbr-2020-0018

Attoe, D. E., & Climie, E. A. (2023). Miss. Diagnosis: A Systematic Review of ADHD in Adult Women. *Journal of Attention Disorders*, *27*(7):645–653. https://doi.org/10.1177/10870547231161533

David, F., Kalibala, G., Pichon, B., & Treur, J. (2024). A Network Model for Modulating Sensory Processing Sensitivity in Autism Spectrum Disorder: Epigenetics, Adaptivity, and Other Factors. *Cognitive Systems Research*, *87*, 101240: 2. https://doi.org/10.1016/j.cogsys.2024.101240

Dosch, A., Rochat, L., Ghisletta, P., Favez, N., & Van Der Linden, M. (2016). Psychological Factors Involved in Sexual Desire, Sexual Activity, and Sexual Satisfaction: A Multi-factorial Perspective. *Archives of Sexual Behavior*, *45*(8): 2030. https://doi.org/10.1007/s10508-014-0467-z

Field, T. (2025). Emotion Dysregulation Research: A Narrative Review. *Journal of Clinical Psychology and Neurology*: 1. https://doi.org/10.61440/JCPN.2025.v3.51

Goldberg, S. Y., Thulin, M. C., Kim, H. S., & Dawson, S. J. (2024). Distressing Problems with Sexual Function and Symptoms of Attention-Deficit/Hyperactivity Disorder. *Archives of Sexual Behavior*, *53*(10): 3739, 3742, 3744. https://doi.org/10.1007/s10508-024-02977-4

Hertz, P. G., Turner, D., Barra, S., Biedermann, L., Retz-Junginger, P., Schöttle, D., & Retz, W. (2022). Sexuality in Adults With ADHD: Results of an Online Survey. *Frontiers in Psychiatry*, *13*, 868278: 1, 5, 11. https://doi.org/10.3389/fpsyt.2022.868278

Herz, R. S., & Cahill, E. D. (1997). Differential use of sensory information in sexual behavior as a function of gender. *Human Nature*, *8*(3): 279–280. https://doi.org/10.1007/BF02912495

Jurek, L., Duchier, A., Gauld, C., Hénault, L., Giroudon, C., Fourneret, P., Cortese, S., & Nourredine, M. (2025). Sensory Processing in Individuals With Attention-Deficit / Hyperactivity Disorder Compared With Control Populations: A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *64*(10): 1132-1133,1140-1145. https://doi.org/10.1016/j.jaac.2025.02.019

Kamath, M. S., Dahm, C. R., Tucker, J. R., Huang-Pollock, C. L., Etter, N. M., & Neely, K. A. (2020). Sensory profiles in adults with and without ADHD. *Research in Developmental Disabilities*, *104*, 103696: 2, 12-13. https://doi.org/10.1016/j.ridd.2020.103696

Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., Howes, M. J., Jin, R., Secnik, K., Spencer, T., Ustun, T. B., & Walters, E. E. (2005). The World Health Organization adult ADHD self-report scale (ASRS): A short screening scale for use in the general population. *Psychological Medicine*, *35*(2): 245–256. https://doi.org/10.1017/S0033291704002892

Laumann, E. O., Paik, A., & Rosen, R. C. (1999). Sexual Dysfunction in the United States: Prevalence and Predictors. *JAMA*, *281(6): 540*. https://doi.org/doi:10.1001/jama.281.6.537

Leiner, D. J. (2021). *SoSci Survey* (Version Version 3.2.31) [Computer software]. https://www.soscisurvey.de

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety & Stress Scales (DASS--21, DASS--42)*. https://doi.org/10.1037/t01004-000

Metz, A. E., Boling, D., DeVore, A., Holladay, H., Liao, J. F., & Vlutch, K. V. (2019). Dunn's Model of Sensory Processing: An Investigation of the Axes of the Four-Quadrant Model in Healthy Adults. *Brain Sciences*, *9*(2), 35: 1-2. https://doi.org/10.3390/brainsci9020035

Ogrodnik, M., Karsan, S., & Heisz, J. J. (2023). Mental Health in Adults With ADHD: Examining the Relationship With Cardiorespiratory Fitness. *Journal of Attention Disorders*, *27*(7): 701. https://doi.org/10.1177/10870547231158383

Panagiotidi, M., Overton, P. G., & Stafford, T. (2018). The relationship between ADHD traits and sensory sensitivity in the general population. *Comprehensive Psychiatry*, *80*: 179–182. https://doi.org/10.1016/j.comppsych.2017.10.008

Pascoal, P. M., Raposo, C. F., & Roberto, M. S. (2020). A Transdiagnostic Approach to Sexual Distress and Sexual Pleasure: A Preliminary Mediation Study with Repetitive Negative Thinking. *International Journal of Environmental Research and Public Health*, *17*(21), 7864: 15387. https://doi.org/10.3390/ijerph17217864

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). *Factor structure of the Barratt Impulsiveness Scale*. *51 (6):* 768-774.

https://doi.org/10.1002/1097-4679(199511)51:6<768::AIDJCLP2270510607>3.0.CO;2-1

Puszcz, A., Platnikow, M., Antos, Z., Czech, M., & Kipka, A. (2025). Sexual Functioning in Individuals With Attention-Deficit Hyperactivity Disorder: A Narrative Review. *Cureus: 1, 4-5*. https://doi.org/10.7759/cureus.97194

Reid, R. C., Garos, S., & Fong, T. (2012). Psychometric development of the hypersexual behavior consequences scale. *Journal of Behavioral Addictions*, *1*(3): 115–122. https://doi.org/10.1556/JBA.1.2012.001

Robertson, A. E., & Simmons, D. (2025, February 28). *The Glasgow Sensory Questionnaire*. Open Science Framework. https://osf.io/h95n4/overview

Rosenberg, M. (1979). *Conceiving the Self* (Vol. 1979). New York: Basic Books.

Rothmore, J. (2020). Antidepressant-induced sexual dysfunction. *Medical Journal of Australia*, *212*(7): 329. https://doi.org/10.5694/mja2.50522

Schulze, M., Aslan, B., Jung, P., Lux, S., & Philipsen, A. (2022). Robust perceptual-load-dependent audiovisual integration in adult ADHD. *European Archives of Psychiatry and Clinical Neuroscience*, *272*(8): 1443-1444, 1446-1447, 1449. https://doi.org/10.1007/s00406-022-01401-z

Soldati, L., Bianchi-Demicheli, F., Schockaert, P., Köhl, J., Bolmont, M., Hasler, R., & Perroud, N. (2020). Sexual Function, Sexual Dysfunctions, and ADHD: A Systematic Literature Review. *The Journal of Sexual Medicine*, *17*(9): 1654, 1660. https://doi.org/10.1016/j.jsxm.2020.03.019

Stulhofer, A., Buško, V., & Brouillard, P. (2011). *The New Sexual Satisfaction Scale and its short form. Handbook of sexuality-related measures:* 530-532.

Wallin, K., Wallin Lundell, I., Hanberger, L., Alehagen, S., & Hultsjö, S. (2022). Self-experienced sexual and reproductive health in young women with Attention Deficit Hyperactivity Disorder: A qualitative interview study. *BMC Women's Health*, *22*(1), 289: 1-2, 6. https://doi.org/10.1186/s12905-022-01867-y

Watt, J. D., & Ewing, J. E. (1996). Toward the development and validation of a measure of sexual boredom. *Journal of Sex Research*, *33*(1): 57–66. https://doi.org/10.1080/00224499609551815

Williams, O. C., Prasad, S., McCrary, A., Jordan, E., Sachdeva, V., Deva, S., Kumar, H., Mehta, J., Neupane, P., & Gupta, A. (2023). Adult attention deficit hyperactivity disorder: A comprehensive review. *Annals of Medicine & Surgery*, *85*(5): 1803–1804. https://doi.org/10.1097/MS9.0000000000000631

Wymbs, B. T., Canu, W. H., Sacchetti, G. M., & Ranson, L. M. (2021). Adult ADHD and romantic relationships: What we know and what we can do to help. *Journal of Marital and Family Therapy*, *47*(3): 666. https://doi.org/10.1111/jmft.12475

Young, S., & Cocallis, K. (2023). A Systematic Review of the Relationship Between Neurodiversity and Psychosexual Functioning in Individuals with Autism Spectrum Disorder (ASD) or Attention-Deficit/Hyperactivity Disorder (ADHD). *Neuropsychiatric Disease and Treatment*, *Volume 19*: 1379. https://doi.org/10.2147/NDT.S319980

Ethical and Psychological Aspects of the Use of Generative AI by Junior Doctors

Karina Zamrazilová

Faculty of Humanities, Charles University, Prague, Czech Republic
karina.zamrazilova@fhs.cuni.cz

Abstract

With the rapid development of generative artificial intelligence (GenAI), its integration into clinical decision-making is becoming an increasingly common component of contemporary medical practice. This pilot study, conducted within the framework of an ongoing doctoral dissertation project, focuses on the psychological and ethical dimensions of GenAI use among junior doctors in the Czech Republic and Germany. The research is based on a qualitative analysis of semi-structured interviews. Preliminary findings suggest emerging forms of psychological dependence on AI tools, particularly in the form of automation bias, compulsive use, anthropomorphization of AI systems, and gradual deskilling. The study further highlights potential manifestations of cognitive delegation, the weakening of critical thinking, and the transformation of decision-making autonomy within the broader context of human–AI interaction.

1 Introduction

In 2022, following the public release of ChatGPT 3.5, many scholars and technology experts anticipated the emergence of a new form of technological revolution, often framed in optimistic terms. It was widely assumed that so-called hybrid teams composed of human and non-human actors would become an integral part of professional environments, particularly within the broader framework of human–AI interaction (Dwivedi et al., 2023). Until early 2025, the majority of published studies emphasized the potential benefits of GenAI in healthcare, especially regarding clinical decision-making and administrative support for physicians (Andrew, 2024). For instance, a study examining the use of generative AI models to reduce physicians' cognitive biases concluded that integrating multi-agent discussions significantly increased diagnostic accuracy (Ke et al., 2024).

However, in 2025, a groundbreaking study conducted at MIT examined the effects of generative AI use on cognition, memory, and neural activity during essay writing tasks. Researchers compared groups using ChatGPT, internet search engines, and only their own knowledge while monitoring brain activity through EEG and analyzing the produced texts. The study concluded that the use of ChatGPT reduced cognitive engagement, memory recall, and the sense of authorship, whereas participants writing without AI demonstrated higher neural activity, better memory retention, and greater originality. The authors further argued that long-term reliance on AI systems may lead to so-called "cognitive debt," characterized by the weakening of critical thinking, independent learning, and deeper information processing, despite the increased efficiency provided by AI tools (Kosmyna et al., 2025). These findings are further supported by another recent study suggesting that physicians may rapidly develop dependence on AI-assisted systems (Budzyń et al., 2025).

This study focuses on the psychological, ethical, and cognitive dimensions of generative AI use among junior doctors, particularly on the transformation of decision-making autonomy, professional responsibility, and trust in technological systems. Special attention is devoted to potential manifestations of cognitive delegation, automation bias, deskilling, and emerging forms of psychological dependence on AI tools. This pilot study is based on a preliminary qualitative analysis of semi-structured interviews and forms part of an ongoing doctoral dissertation project.

2 Related Work

According to the World Health Organization (WHO), the global healthcare system is expected to face a shortage of 18 million healthcare workers by 2030. Current research therefore emphasizes the urgent need to address this situation in order to prevent healthcare professionals from leaving the field. Ye et al. (2022) further argue that junior doctors in the early stages of their careers represent the most vulnerable group in terms of completely abandoning the medical profession.

A junior doctor may be defined as an individual who has successfully completed medical school but must continue postgraduate clinical training in order to obtain board certification in a medical specialization (Verywell Health, 2023). Another defining characteristic is that junior doctors work under the supervision of a fully certified senior physician who should, in principle, be available at all times to provide assistance or consultation when necessary (MyHSN, 2026). However,

previous research suggests that such supervision is often insufficient or unavailable in practice (Zamrazilová, 2023). The healthcare systems examined in this study, namely those of the Czech Republic and Germany, are both structured according to this principle.

Gonzalez and Heidari (2025) focused on the distinction between static and dynamic decision-making, with the medical profession clearly belonging to the latter category. Data-driven AI systems may therefore function as tools that provide recommendations and guidance to individuals responsible for making final clinical decisions. However, a major limitation of current AI systems lies in their difficulty adapting to dynamic environments by incorporating real-time human input and automatically adjusting recommendations in response to changing goals, shifting contexts, user feedback, and evolving environmental conditions.

Another critical issue is that individuals frequently fail to appropriately calibrate their reliance on AI during decision-making processes. The authors argue that excessive dependence on GenAI outputs without sufficient critical evaluation may result in suboptimal decisions. One of the most common cognitive biases in this context is anchoring bias, which occurs when initial information exerts a disproportionate influence on subsequent judgments and decisions. This effect persists even when the anchor itself is arbitrary or irrelevant. Anchoring introduces a cognitive reference frame that subsequently shapes an individual's final decisions (Carter & Liu, 2025).

Language also represents a significant challenge for the use of GenAI in medicine, as language is not merely a communication tool but also a framework shaping thought, perception of reality, and behavior. In medical contexts, language structures diagnoses, treatment decisions, risk communication, and patient trust. The choice of words and metaphors (e.g., "mechanical dysfunction," "war," or "battle") may influence whether physicians lean toward aggressive treatment, palliative care, or pharmacological intervention. Cultural and linguistic differences therefore play an important role in medicine, and misunderstandings in communication may contribute to misdiagnosis (Kumar & Arora, 2025). This issue is further complicated by the fact that most GenAI systems have primarily been trained on English-language medical datasets (Andrew, 2024), potentially contributing to additional forms of cognitive bias.

As previously noted, GenAI systems are capable of analyzing medical records, patient narratives, and clinical notes while assisting with diagnosis (Kumar et al., 2025). Initially, GenAI was widely expected to function as an intermediary capable of reducing cognitive biases by providing objective and comprehensive information. This optimism reflected the belief that AI-assisted systems could surpass human judgment in terms of fairness, accuracy, and efficiency. However, emerging evidence now suggests that GenAI may instead amplify existing cognitive biases rather than mitigate them (Rodriguez-Fernandez, 2026).

3 Methodology

For this qualitative study, a methodology based on semi-structured online and in-person interviews was employed. Each participant took part in three to five in-depth interviews conducted between 2023 and the present.

3.1 Sample

The current sample consists of 16 participants, of whom 9 are from the Czech Republic and the remaining 7 from Germany. All participants are junior doctors between the ages of 26 and 34. The sample is evenly divided by gender, comprising 8 men and 8 women. Participants were recruited through snowball sampling and via announcements posted on social media platforms, through which individual participants subsequently volunteered for the study. An important finding is that all participants reported experiencing mental health difficulties and regularly attending psychotherapy, while some also reported receiving psychiatric treatment. All participants associated their psychological distress with the medical profession and the working conditions they experience. All participants included in this study were strictly anonymized.

4 Analysis and Interpretation

A qualitative thematic analysis was employed to analyze the data obtained from the semi-structured interviews, with the aim of identifying recurring patterns of meaning and dominant themes within the participants' narratives. The analysis focused on how individual actors describe and conceptualize generative artificial intelligence, the purposes for which they use these systems, the nature of their relationship to AI technologies, what GenAI personally represents to them, and the role it plays in their everyday clinical practice. Particular attention was devoted to themes related to decision-making autonomy, trust in technological systems, psychological dependence on AI, and transformations of clinical judgment within the broader context of human–AI interaction.

Preliminary findings suggest that 11 out of 16 participants exhibit signs of psychological dependence on these technologies, particularly in the form of automation bias, compulsive use in both personal and professional contexts, and the anthropomorphization of AI systems. Another recurring phenomenon identified in the interviews was so-called deskilling, understood as the gradual weakening of professional com-

petencies due to increasing reliance on AI-assisted systems. The findings further indicate that all participants routinely incorporate these technologies into their daily clinical practice. Differences between participants from the Czech Republic and Germany appear to be negligible. The findings suggest that these emerging technologies are significantly transforming decision-making autonomy, perceptions of professional responsibility, and trust in technological systems. Generative AI is simultaneously reshaping the ways in which individuals approach their own judgment, decision-making processes, and everyday professional activities. Such developments may contribute not only to progressive cognitive delegation to AI systems, but also to the emergence of new forms of psychological dependence on these technologies.

5 Discussion

The interaction between humans and generative artificial intelligence represents a specific type of relationship in which no independent human actor exists beyond the user and the technological system itself. This form of interaction may contribute to risks of self-confirmation, artificial consensus, and the gradual weakening of critical reflection, as AI systems frequently reproduce or reinforce users' existing attitudes, assumptions, and patterns of thinking (Silkej & Fredriksson, 2020). Within a broader societal context, these processes raise important questions concerning the impact of AI on democratic decision-making, individual autonomy, and potential forms of digital cognitive decline. At the same time, prolonged interaction with AI tools appears to influence how individuals approach personal judgment, professional responsibility, and decision-making processes.

This study represents a pilot qualitative investigation based on a limited number of participants; therefore, its findings cannot be generalized to the wider population. The primary aim of the study is to open a broader discussion concerning the psychological and social implications of generative AI within medical environments and to identify key areas for future research. A deeper understanding of these phenomena will require more extensive qualitative and quantitative studies focused on the long-term impact of AI on clinical decision-making, professional autonomy, and the psychological dimensions of human–technology interaction.

6 Conclusion

Empirical data from this pilot qualitative study suggest that the use of generative artificial intelligence in medical settings may lead not only to changes in clinical decision-making, but also to transformations in individuals' relationships to their own judgment, professional responsibility, and everyday clinical practice. A significant proportion of participants exhibited signs of psychological dependence on AI tools, particularly in the form of automation bias, compulsive use, anthropomorphization of AI systems, and gradual deskilling. The findings further indicate that, for many junior doctors, GenAI has become an integral part of everyday clinical reality, influencing not only work efficiency but also patterns of thinking and decision-making.

The results support theoretical concepts related to cognitive delegation, technological dependence, and the transformation of autonomy within the context of human–AI interaction. At the same time, the study raises broader questions concerning the long-term impact of AI on the mental health of healthcare professionals, physicians' professional identity, and the ethical dimensions of algorithmically assisted decision-making. Although the findings cannot be generalized due to the limited sample size, the study highlights the need for further interdisciplinary research integrating psychology, medicine, anthropology, and AI studies. Future research should focus particularly on the long-term effects of GenAI on clinical judgment, critical thinking, professional autonomy, and potential forms of digital cognitive decline.

Acknowledgement

This work was supported by institutional resources of Faculty of Humanities, Charles University, Prague.

References

Budzyń, K. et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903.

Carter, L. and Liu, D. (2025). How was my performance? exploring the role of anchoring bias in ai-assisted decision making. *International Journal of Information Management*, 82:102875.

Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). So what if chatgpt wrote it? multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71.

Gonzalez, C. and Heidari, H. (2025). A cognitive approach to human–ai complementarity in dynamic decision-making. *Nature Reviews Psychology*, pages 1–15.

Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., and Ting, D. S. W. (2024). Development and testing of retrieval augmented generation in large language models—a case study report. *arXiv*.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., and Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *MIT Media Lab*. Preprint.

Kumar, A. and Arora, M. (2025). Cognitive linguistics in medicine: How language shapes perception, decision-making and patient care. *Cognitive Linguistics*, 2(02).

MyHSN (2026). What is a resident doctor? Available online: https://www.myhsn.co.uk/top-tip/what-is-a-resident-doctor/.

Rodriguez-Fernandez, F. (2026). Artificial intelligence and economic psychology: toward a theory of algorithmic cognitive influence. *AI & SOCIETY*, 41(2):1481–1492.

Silkej, E. (2020). Linguistic differences in real conversations: Human to human vs human to chatbot. Student project report, University of Gothenburg, Gothenburg, Sweden.

Verywell Health (2023). Types of doctors: Residents, interns, and fellows. Available online: https://www.verywellhealth.com/types-of-doctors-residents-interns-and-fellows-3157293.

Ye, X., Zhai, M., Feng, L., Xie, A., Wang, W., and Wu, H. (2022). Still want to be a doctor? medical student dropout in the era of covid-19. *Journal of Economic Behavior & Organization*, 195:122–139.

Zamrazilová, K. (2023). Nás soudila nemoc/kyslík, prosím: Reflexe povinné služby mladých doktorů a mediků během covidu-19. Master's thesis, Charles University, Prague, Czech Republic.

Large Behavior Models: a promising approach for safe and reliable robotics?

Branislav Zigo, Igor Farkaš

Department of Applied Informatics, Comenius University Bratislava
{branislav.zigo, igor.farkas}@fmph.uniba.sk

Abstract

Vision-Language-Action models (VLAs) and Large Behavior Models (LBMs) are becoming a central research direction for cognitive and humanoid robotics. They connect semantic perception, natural-language task specification, multimodal state estimation, and continuous motor control in one trainable policy stack. This paper reviews the motivation, historical roots, architectural principles, and taxonomy of recent VLA and LBM approaches. Particular attention is given to multimodal and multisensory fusion, cross-embedding of vision, language, robot state and action, autoregressive action token generation, diffusion policy, and flow matching. The paper defines safety and reliability as operational properties of embodied policies: constraint satisfaction, calibrated uncertainty, recoverability from failures, robustness under distribution shift, and transparent human interaction. The main conclusion is that cognitive robotics is moving toward hybrid architectures that combine pretrained semantic backbones, multisensory latent state estimation, generative continuous action heads and predictive world models. In the end, we also discuss the limitations of this approach to building intelligent robotic systems. With reference to the title, LBMs are promising substrates for safe and reliable robotics only when embedded in monitored, uncertainty-aware and verifiable robot-control stacks.

1 Introduction

Humanoid robots are no longer only mechanical platforms for scripted motion. They are becoming embodied AI systems that must interpret human instructions, perceive complex scenes, interact safely with objects and people, and coordinate many degrees of freedom in real time. This requirement creates a difficult cognitive-robotics problem: the robot must connect symbolic or linguistic task descriptions with perceptual grounding, physical state estimation, and closed-loop control. Vision-Language-Action models (VLAs) address this problem by extending Vision-Language Models (VLMs) toward direct robot action generation (Kawaharazuka et al., 2025). Large Behavior Models (LBMs) address a broader problem: how to scale robot policies across many behaviors, tasks, environments and embodiments (TRI LBM Team et al., 2026).

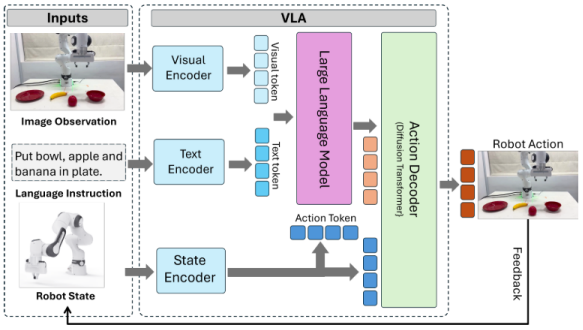


Fig. 1: Architecture of a typical Vision-Language-Action model for robotic manipulation (Din et al., 2025). The three embeddings are fused in an LLM that generates multimodal semantic representation of the intended task, i.e., along with robot state features, decoded to a trajectory that accomplishes the task.

The distinction between VLA and LBM is useful but not absolute. A VLA is usually language-grounded: it receives an image or video observation, a language instruction, and possibly proprioceptive state, then produces actions (see an example in Fig.1). An LBM is a large-scale behavior policy trained on broad robot data; it may be language-conditioned, but it can also be defined primarily by action competence and data scale. The strongest current robot foundation models increasingly combine semantic grounding from pretrained vision-language backbones, broad behavioral competence from large trajectory datasets, and continuous action generation for physical control (Octo Model Team et al., 2024; Black et al., 2024; NVIDIA et al., 2025).

For cognitive and humanoid robotics, this shift is important because classical modular pipelines often separate perception, mapping, planning, grasp synthesis and motor control. Such modularity remains valuable for verification and safety, but it can be brittle under open-ended instructions, clutter, occlusion, deformable objects, contact-rich manipulation and embodiment mismatch. VLA and LBM research does not remove the need for structure; rather, it asks how much of this structure can be learned, shared across tasks and adapted to new robots.

There also exist recent robotics-related proto-LBMs that stand somewhere between VLAs and LBMs. RoboVLM (Li et al., 2026) and HY-Embodied (Yu

et al., 2026) differ in details but converge on the same idea: a unified, scalable VLA foundation model that generates behavior directly, an emerging form of LBMs.

The contribution of the paper is threefold: (1) it surveys current VLA/LBM architectures, (2) synthesizes their trade-offs for robotic use, and (3) proposes minimal safety-and-reliability evaluation criteria that should accompany any claim of progress toward safe humanoid deployment.

2 Historical roots and motivation

The first root of VLAs and LBMs is end-to-end visuomotor learning. Early deep visuomotor policies showed that perception and control could be trained jointly from raw sensory observations, challenging the idea that a robot always needs a manually specified symbolic state before action selection (Noda et al., 2014). RT-1 later demonstrated that transformer architectures can absorb large real-world robot datasets and generalize across many tasks when camera observations, instructions and motor commands are represented in a compact sequence format (Brohan et al., 2022).

The second root is multimodal representation learning that studies how models represent, align, translate, fuse and co-learn from multiple modalities (Baltrušaitis et al., 2019). Early deep multimodal learning showed that coordinated representations can transfer information across sensory streams (Ngiam et al., 2011). In robotics, multimodal integration was already studied on humanoid platforms by learning fused sensorymotor time-series representations for behavior generation (Noda et al., 2014). This tradition becomes essential in manipulation, where vision may see the object globally, proprioception tells the robot where its body is, and tactile or force feedback reveals contact that cameras cannot observe (Li et al., 2023; Sferrazza et al., 2024).

The third root is language-conditioned imitation learning and vision-language pretraining. CLIP showed that large-scale image-text pretraining can produce transferable visual concepts addressable through language (Radford et al., 2021). CLIPort then combined CLIP-like semantic understanding with spatial manipulation precision, separating the question of what to manipulate from where to act (Shridhar et al., 2022). PaLM-E inserted continuous embodied observations into the token stream of a large language model (LLM), showing that language, image and robot state can be processed jointly for embodied reasoning (Driess et al., 2023). VIMA generalized the prompting view by interleaving visual and textual prompt tokens and decoding robot actions autoregressively (Jiang et al., 2022).

The final root is sequence modelling for decision making. Decision Transformer and Trajectory Transformer reframed control as sequence prediction over states, actions and returns (Chen et al., 2021; Janner et al., 2021). This was conceptually important because it made robot control compatible with the scaling recipes of LLMs: tokenize inputs, train on large sequential datasets, and predict the next element in context. RT-2 made this connection explicit by expressing robot actions as text-like tokens and co-training on robot trajectories and web-scale vision-language tasks (Zitkovich et al., 2023).

3 Key architectural principles

3.1 Multimodal and multisensory fusion

Multimodal fusion means that a policy does not rely on one input channel only. In VLAs and LBMs, the usual modalities are image or video, language instruction, robot proprioception, action history and sometimes tactile, force, audio or depth observations (Baltrušaitis et al., 2019). Multisensory fusion is especially important for humanoid manipulation because many task-relevant variables are hidden: the object may be occluded by the hand, contact pressure may change before the object visibly moves, and slippage may be felt before it is seen. Experiments on visual-auditory-tactile manipulation show that different senses dominate different parts of the task: vision supplies global scene context, audio can mark contact events, and touch supplies local geometry and contact information (Li et al., 2023). Vision-touch masked multimodal learning similarly shows that tactile information improves sample efficiency and robustness under occlusion (Sferrazza et al., 2024).

For cognitive robotics, the key point is that the robot internal state should not be a simple image caption. It should be a control-relevant latent state: a compact representation that binds object identity, spatial relations, affordances, body configuration, contact state and task intention. Here the cross-embedding becomes central.

3.2 Cross-embedding and embodied tokens

Cross-embedding refers to representing multiple modalities in a common or coordinated latent space so that information can be compared, attended to, fused or translated across modalities (Ngiam et al., 2011; Baltrušaitis et al., 2019). In current robot foundation models, this often means mapping visual features, language tokens and robot state vectors into a transformer-compatible sequence. PaLM-E is a clear example: visual observations and continuous state estimates are projected into the same processing stream as language tokens (Driess et al., 2023). VIMA uses multimodal prompts in which text and visual tokens specify the manipulation task (Jiang et al., 2022). OpenVLA combines a Llama-based decoder with fused DINOv2 and SigLIP visual features, illustrating how pretrained visual and vision-language

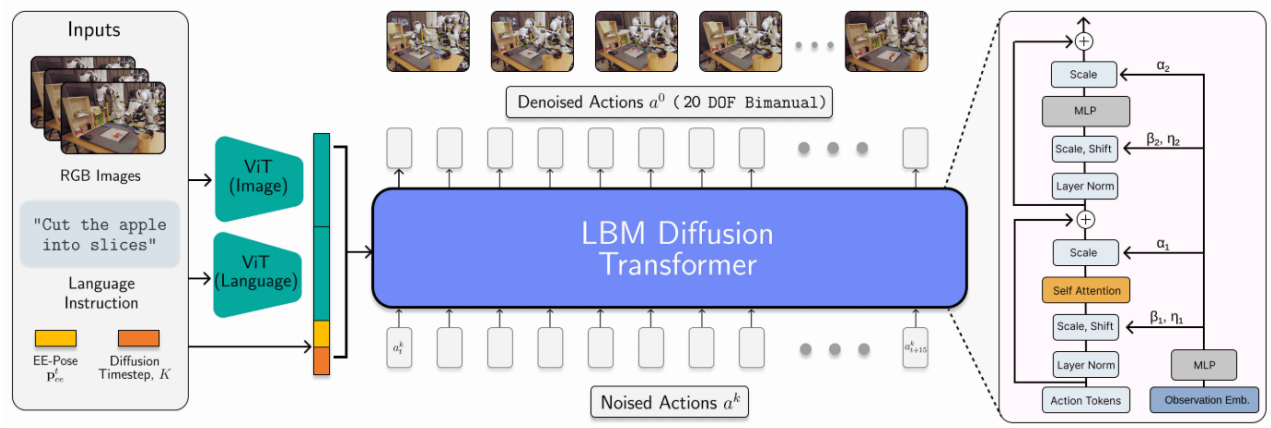


Fig. 2: Architecture of a Large Behavior Model (TRI LBM Team et al., 2026) The model is conditioned on language, vision and proprioception and outputs 20-dimensional actions over 16 timesteps.

encoders can support robot action prediction (Kim et al., 2025; Oquab et al., 2023; Zhai et al., 2023).

This design has a cognitive interpretation. The robot does not merely classify pixels or parse commands. It constructs an embodied context in which words such as cup, fragile, left, gently or give become linked to visual objects, body posture, force constraints and possible action sequences. Cross-embedding is therefore a computational mechanism for grounding meaning in action.

3.3 Autoregressive token generation

Autoregressive token generation means that the model predicts the next token conditioned on previous tokens. In language models the tokens are words or subwords; in robot policies they may be discretized actions, action chunks or symbolic representations of continuous control variables (Chen et al., 2021). RT-2 is the canonical VLA example: it formats robot actions as text-like tokens so that the same model can learn both language responses and action outputs (Zitkovich et al., 2023). OpenVLA keeps this transformer-based VLA idea but releases an open model and training stack, making the approach more reproducible (Kim et al., 2025).

The advantage of autoregressive VLAs is conceptual and engineering simplicity. They fit naturally with pretrained VLMs and allow web-scale semantic knowledge to influence robot control. Their limitation is action precision. Fine motor control often requires smooth, high-frequency, continuous trajectories. Discretizing a dexterous hand, wrist and arm into tokens can introduce quantization error, longer sequences and latency. This limitation motivates diffusion and flow-based action heads.

4 VLA and LBM model families

A practical taxonomy can be organized by the way task context and action are represented. The categories below should not be read as mutually exclusive boxes. They are design choices along several axes: semantic grounding, temporal horizon, action representation, data scale, embodiment coverage, online feedback and explicit safety mechanisms. This comparative reading is more useful for robotics than a simple chronological list because the most suitable model family depends on the task regime.

(1) Semantic-spatial manipulation models use vision-language pretraining to ground object concepts while preserving precise spatial control. CLIPort is the representative example because it combines a semantic pathway for what to manipulate with a spatial pathway for where to manipulate (Shridhar et al., 2022). Such models are not always full VLAs in the recent foundation-model sense, but they are historically important because they demonstrated that internet-scale semantic representations can improve robot manipulation.

(2) Promptable and embodied language models process text, visual observations and sometimes robot state in a common sequence. PaLM-E is primarily an embodied multimodal language model for reasoning and planning (Driess et al., 2023), whereas VIMA is a transformer agent that solves tabletop tasks from multimodal prompts and outputs actions autoregressively (Jiang et al., 2022). These models make the instruction interface explicit and show how language can structure robot behavior beyond fixed task labels.

(3) Autoregressive VLA controllers directly treat action as part of the token sequence. RT-1 introduced a scalable robotics transformer trained on diverse real-world robot data (Brohan et al., 2022). RT-2 then transferred web-scale vision-language knowledge into robot control by co-training on internet vision-language tasks and robot trajectories (Zitkovich et al., 2023). Open-

VLA extended this line as an open 7B-parameter VLA trained on 970k robot demonstrations from Open X-Embodiment (Kim et al., 2025; O'Neill et al., 2023).

(4) Generalist robot policies and LBMs emphasize behavioral scale across tasks and embodiments. Octo is a transformer-based diffusion policy trained on 800k trajectories from Open X-Embodiment and designed for adaptation to new sensors and action spaces (Octo Model Team et al., 2024; O'Neill et al., 2023). TRI's LBMs extend the diffusion-policy paradigm across approximately 1,700 hours of robot data and evaluate multitask dexterous manipulation with controlled real-world and simulation trials (TRI LBM Team et al., 2026). In this family, language is useful, but the defining feature is broad behavioral competence from large robot data.

(5) Continuous generative policies use diffusion or flow matching to produce smooth action trajectories. Diffusion Policy models actions as a denoising process conditioned on observations (Chi et al., 2025). RDT-1B scales a diffusion transformer for bimanual manipulation (Liu et al., 2024). GR00T N1 uses a dual-system VLA architecture for humanoids in which a vision-language module interprets the instruction and a diffusion transformer generates motor actions (NVIDIA et al., 2025). The π_0 model uses flow matching rather than classical action tokenization to generate continuous actions from a pretrained vision-language backbone (Black et al., 2024).

(6) Predictive world-model alternatives such as JEPA-based systems learn how the world evolves in latent space and then plan actions, instead of directly mapping observation and instruction to action (Assran et al., 2025; Terver et al., 2025). They are not merely another action head; they represent a different hypothesis about cognitive robotics: robust behavior may require prediction and planning in an abstract world model. The key trade-off is therefore not "which family is best" but "which failure mode is acceptable for a given robot task." Autoregressive token policies are attractive when instruction following and semantic transfer are central, but they may suffer from discretization and latency. Diffusion and flow-matching policies are better matched to smooth dexterous trajectories, but they require careful sampling budgets and closed-loop replanning. JEPA-style world models are appealing for lookahead and causal anticipation, but their language-conditioned action interfaces are less mature. Safety-critical humanoid applications will probably require combinations of these mechanisms rather than a single dominant paradigm.

5 Datasets and scaling

The scaling of VLAs and LBMs depends on datasets as much as architectures. Open X-Embodiment pooled robot trajectories across many institutions and robot embodiments, enabling the RT-X line of cross-robot policies and providing a foundation for OpenVLA and Octo (O'Neill et al., 2023; Kim et al., 2025; Octo Model Team et al., 2024). DROID extended the dataset landscape with in-the-wild robot manipulation demonstrations collected across many real scenes, tasks and data collectors (Khazatsky et al., 2024). VIMA-Bench, introduced with VIMA, provides procedurally generated tabletop tasks with multimodal prompts, supporting systematic evaluation of generalization (Jiang et al., 2022). Another interesting dataset is Humanoid Everyday that aims to advance research in general-purpose humanoid manipulation and lay the groundwork for more capable and embodied robotic agents in real-world scenarios (Zhao et al., 2025).

However, robot data remain scarce compared with internet text and image data. This is the central bottleneck for humanoid robotics. Humanoids have high-dimensional bodies, many coupled joints, unstable contacts and safety constraints. Consequently, near-future VLA and LBM training will likely combine robot teleoperation, simulation, motion capture, human video, synthetic data and self-supervised world-model pretraining (O'Neill et al., 2023; Khazatsky et al., 2024; NVIDIA et al., 2025; Assran et al., 2025).

From the safety perspective, data scale alone is insufficient. Datasets should explicitly contain recovery behavior, near misses, ambiguous instructions, sensor dropouts, out-of-distribution objects and human-interaction cases. Otherwise, a policy may learn impressive average-task performance while remaining poorly calibrated in precisely the cases where a humanoid robot must slow down, ask for clarification or retreat.

6 Diffusion policy and flow matching

6.1 Diffusion policy

Diffusion policy applies the idea of denoising diffusion models to robot action generation (Chi et al., 2025). During training, the model sees expert action trajectories, adds different levels of noise to them, and learns how to remove that noise while conditioning on observations such as images, robot state and task context. During inference, it begins with a noisy action sequence and iteratively denoises it into a plausible action trajectory. This is useful because manipulation is often multimodal: there may be several correct ways to grasp a cup, fold a cloth or move around an obstacle. A standard regression policy may average these alternatives into a bad action, while a diffusion policy can represent several possible action modes (Chi et al., 2025).

In robotics, diffusion policies are often used with receding-horizon control (Chi et al., 2025). The policy generates an action chunk, the robot executes the first part, receives updated observations, and then replans. This provides a compromise between long-horizon tem-

poral consistency and closed-loop correction. Octo, RDT-1B, GR00T N1 and TRI LBMs can all be read as different ways of scaling this continuous generative-action idea to larger datasets, broader tasks or embodiments (Octo Model Team et al., 2024; Liu et al., 2024; NVIDIA et al., 2025; TRI LBM Team et al., 2026).

The price of diffusion is sampling cost. Since the action is refined through multiple denoising steps, inference may be slower than one-shot prediction. This matters for dexterous hands, whole-body humanoids and dynamic interaction. Research therefore explores fewer denoising steps, action chunking, specialized diffusion transformers and alternatives such as flow matching (Lipman et al., 2023; Black et al., 2024).

6.2 Flow matching and the velocity-field analogy

Flow matching is a generative modelling objective that learns a vector field transforming simple noise into samples from the data distribution (Lipman et al., 2023). The intuitive analogy is a river current. Imagine many small floating leaves initially scattered randomly on the water. A velocity field tells each leaf in which direction and how fast to move at each point. If the current is learned correctly, the random cloud of leaves gradually flows into the shape of the desired object. In a robot policy, the random cloud is initial action noise, the final shape is a meaningful action trajectory, and the learned current is the policy's velocity field (Black et al., 2024).

Compared with classical diffusion, flow matching can be understood as learning a more direct transport path from noise to data. It does not only ask the model to remove noise step by step; it asks the model to learn the instantaneous direction of travel along a probability path. The π_0 model uses this principle to generate continuous robot actions with a separate action expert on top of a pretrained vision-language backbone (Black et al., 2024). This is attractive for fine motor control because actions remain continuous, action chunks can be smooth, and the semantic backbone can still provide language and visual grounding.

For an audience new to the velocity-field concept, a second analogy is navigation in a city. A map tells where the streets are, but a velocity field tells how traffic should move at every location and time. Flow matching trains the policy to produce this motion rule. Once trained, the model can start from an unstructured initial action and follow the learned flow toward a coherent motor command. For humanoid robotics, this is promising because the same high-level instruction may require subtle continuous coordination of arm parts (i.e. shoulder, elbow, etc.).

7 JEPA and predictive world-models

Joint-Embedding Predictive Architectures (JEPAs) offer a different route toward embodied intelligence. A JEPA predicts future or missing representations in latent space rather than reconstructing raw pixels (LeCun, 2022; Assran et al., 2023). I-JEPA demonstrated this principle for images (Assran et al., 2023), while V-JEPA 2 extends it to video and physical planning by learning from large-scale video and then adding action-conditioned latent dynamics for robotic tasks (Assran et al., 2025). JEPA-based world models aim to support planning by predicting how the world will change under possible actions, and recent work evaluates which architectural and training choices are important for physical planning in latent space (Terver et al., 2025).

The difference from direct VLA policy learning is fundamental. A VLA often maps observation and instruction directly to action. A JEPA-style world model first learns a predictive representation of the world, then uses planning to choose actions that move the latent state toward a goal (LeCun, 2022; Assran et al., 2025). This may be advantageous for long-horizon cognition, causal anticipation and learning from passive video, but it can add planning cost and currently has less mature language-to-action grounding than VLA controllers. A plausible future is hybrid: a VLA proposes grounded actions, while a predictive world model evaluates consequences and supports lookahead.

8 Implications for cognitive and humanoid robotics

The VLA-to-LBM transition changes how we can think about robot cognition. Cognition is no longer located only in symbolic planning or explicit world models. It is distributed across pretrained perceptual backbones, cross-modal embeddings, temporal sequence models, action generators and feedback loops. In humanoid robots, this matters because behavior depends on the whole body: gaze, torso, arms, hands, tactile contact and balance must cooperate. GR00T N1 and related humanoid foundation models explicitly target this whole-body setting (NVIDIA et al., 2025), while bimanual diffusion models such as RDT-1B target the coordination problem of two arms (Liu et al., 2024).

At the same time, several challenges remain open. (1) Generalization is still limited by data coverage: a model can only generalize robustly if its training distribution contains enough variation in objects, scenes, embodiments and failures (O'Neill et al., 2023; Khazatsky et al., 2024; TRI LBM Team et al., 2026).
(2) Evaluation is difficult because robot success is stochastic, hardware-dependent and sensitive to small environmental changes (TRI LBM Team et al., 2026).
(3) Safety and interpretability are unresolved. In human

environments, a humanoid should not only act successfully but also make its intentions understandable, detect uncertainty, and fail safely.
(4) Multimodal fusion must handle missing or unreliable sensors; a camera may be occluded, a tactile sensor may drift, and a language instruction may be ambiguous (Li et al., 2023; Sferrazza et al., 2024).

These limitations suggest that future cognitive humanoid systems will not be pure end-to-end black boxes. They will probably combine learned policies with constraints, safety monitors, symbolic task structure, uncertainty estimation, human feedback and predictive world models. The contribution of VLA and LBM research is that it supplies a scalable learned substrate on which such structured cognition can be built.

9 Safety, reliability and evaluation of LBMs in robotics

To deal with this issue, safety and reliability must be treated as measurable system properties, not as general impressions. In embodied robotics, safety means that the robot respects physical, social and task constraints under nominal and off-nominal conditions. Reliability means that the robot maintains acceptable performance across repeated trials, changed objects, changed scenes, sensor degradation, embodiment differences and ambiguous instructions.

A minimal evaluation protocol should therefore report more than task success. It should include several criteria as proposed in Table 1.These criteria do not replace standard benchmarks, but they make explicit what must be demonstrated before claims about safe and reliable humanoid deployment are credible.

10 Core problems of MLLMs

In general, despite their interesting performance, MLLMs are subject to several shortcomings (which apply to VLAs and LBMs as well). One is the well-known problem of **"hallucinations"** (Ji et al., 2023), which was first identified in LLMs, primarily linked to the symbol grounding problem (SGP) Huang et al. (2023). Since its discovery, there has been a growing number of literature referring to this undesirable phenomenon manifesting itself in various forms (Alansari and Luqman, 2026). Kalai and Vempala (2024) show that, under calibration assumptions and for certain classes of arbitrary facts, hallucination must occur, i.e., it has a statistical lower bound rather than being only an engineering defect. Regarding models' internal representations, the existence of the so-called H-neurons responsible for hallucinations has been reported (Gao et al., 2025). From the psychological perspective, the hallucinating LLMs have been shown to cause delusional spiraling even in very rational users (Chandra et al., 2026). Obviously, the existence of ubiquitous hallucinations poses substantial obstacles to practical deployment of LLMs and raises concerns regarding their reliability in real-world applications.

However, the expectations that SGP can be solved by grounding symbols in robot experience using MLLMs have not been met, since the hallucinations have been shown not to be language-specific (Sun et al., 2024). According to most recent literature, hallucinations have also been found in MLLMs Huang et al. (2025). Vision-Language Models (VLM) often exhibit cross-modal inconsistency, i.e. a discrepancy between an image content and text response. For instance, in case of object-related hallucinations (a frequent case), three types have been identified: object category, object attribute, and object relation. Hallucinating MLLMs are a problem which implies that merely adding modalities to language, in order to address the SGP and achieve understanding, does not automatically provide a solution. The core of the problem must lie elsewhere. For instance, Farkaš et al. (2025) argue that the problem may be due to a fundamentally different (non-developmental) way how world knowledge is acquired in MLLMs compared to humans, the main argument being that in MLLMs the language is central and that they lack non-linguistic world model (unlike people).

Apart from hallucinations, another problem of MLLMs has been identified, labeled **"mirage effect"** as evidence of an illusion of visual understanding in VLMs (Asadi et al., 2026). The authors define mirage as an AI model describing visual features in the absence of any image input. The output of a model affected by the mirage effect, seen in isolation, is indistinguishable from that of normal visual reasoning. The model does this without expressing any uncertainty, lack of confidence, or acknowledging an assumption or hypothetical scenario. The authors explain that contrary to hallucinations, the mirage effect does not necessarily involve inconsistencies or false responses. The response in mirage-mode can be correct in every sense, accompanied by a meticulous reasoning trace, and completely coherent. The main characteristic of the mirage effect, however, is the construction of a false epistemic frame that is not grounded in the provided input. Mirage effect is hence also a problem that hinders the practical deployment of such VLM models.

For robotics, hallucination and mirage-like behavior are not only epistemic errors; they can become physical risks. A VLA that infers a non-existent object, overstates visual certainty, or fills missing tactile information with a plausible but false latent state may initiate an unsafe action. This is why multimodal grounding must be coupled to uncertainty estimation, sensor-consistency checks and permission to stop or ask for clarification.

Tab. 1: Proposed safety-and-reliability checklist for VLA/LBM-based humanoid systems.

Criterion	What to measure	Why it matters for safe robotics
Task performance	Success rate, completion time, number of intervention requests	Separates useful competence from purely plausible behavior
Constraint satisfaction	Collisions, force/torque limit violations, joint-limit violations, unsafe human proximity	Prevents successful-looking policies from being accepted despite unsafe execution
Robustness	Performance under changed objects, lighting, clutter, sensor noise, occlusion and embodiment variation	Tests whether the policy generalizes beyond the training distribution
Uncertainty and abstention	Confidence calibration, clarification requests, safe stopping and refusal to act under ambiguity	Converts uncertainty into safe behavior instead of hallucinated action
Recovery and fail-safe behavior	Ability to retreat, regrasp, replan, stop smoothly and preserve balance after failures	Measures whether failure is bounded rather than catastrophic
Human interpretability	Explanations of intention, predicted motion and reason for failure or refusal	Supports trust calibration and supervision in human environments
Latency and closed-loop stability	Control delay, action update rate, oscillation or instability during contact	Determines whether generative policies can be used in real physical interaction

11 Challenges for LBMs in robotics

Despite growing excitement about the potential of leveraging AI in robotics, it must be acknowledged perceving and acting in the physical world pose greater and different challenges for AI than analysing data in isolation (Billard et al., 2025). More concretely, we cannot expect approaches that excel in purely data and software-based gaming environments, or image or text generation, to be readily applied to real-time sensing, planning, control and navigation for physical machines operating in complex, unpredictable environments, especially those involving human interaction. As a roadmap for AI in robotics, several short-term challenges have been identified (Billard et al., 2025): (1) Improved datasets (for manipulation and navigation tasks), for different tasks and different embodiments and interaction with humans, (2) Techniques to overcome the sim-to-real gap, (3) Combination of model-based and model-free learning, and (4) Generative models for robotics including the use of existing LLMs for semantic inference, development of new large-scale visual models and VLA models for different tasks.

The paper also mentions the long-term challenges, namely (1) life-long learning which would allow robots to acquire new knowledge and skills during their entire operational life, and (2) transfer learning which would enable transferring learned skills to new tasks, different robots and environments, perceptions between robots, or from simulation to physical robots. Given the expanding literature on VLAs, transfer learning is already being addressed intensively, supported by a growing number of large multimodal datasets.

12 Near-future outlook

In the near future, four developments seem likely.

(1) Continuous action generation will become more important for dexterous and humanoid robotics. Autoregressive token policies will remain attractive for semantic grounding, but diffusion and flow matching better match the smooth control needs of hands and whole-body motion (Chi et al., 2025; Black et al., 2024).

(2) Multisensory fusion will move from an optional add-on to a core requirement, especially for manipulation under occlusion, deformable objects and contact-rich tasks (Li et al., 2023; Sferrazza et al., 2024).

(3) Data mixtures will become increasingly heterogeneous. Robot-only data are too expensive, so training will combine robot trajectories, human demonstrations, egocentric video, simulation, synthetic data and self-supervised video pretraining (O'Neill et al., 2023; Khazatsky et al., 2024; Assran et al., 2025).

(4) The boundary between policy models and world models will blur. VLAs will gain predictive modules, and JEPA-style models will gain stronger language grounding and action interfaces (Assran et al., 2025; Terver et al., 2025).

For cognitive robotics, the most promising direction is therefore not a single architecture but a layered recipe: pretrained semantic perception, multisensory state embedding, generative continuous action, predictive latent planning, and safety-aware human interaction. This recipe can support robots that are not only more capable, but also more adaptive and understandable in everyday human environments.

13 Conclusion

VLAs and LBMs represent a shift from task-specific robotic programs toward scalable embodied policies. VLAs emphasize the grounding of language and vision in action, while LBMs emphasize broad behavioral competence across tasks and data. Their foundations lie in multimodal learning, language-conditioned imitation learning, sequence modelling and robot dataset scaling. Current state-of-the-art systems differ in how they represent actions: autoregressive token models connect

naturally to language-model training; diffusion policies represent multimodal continuous trajectories; flow-matching policies learn a velocity field from noise to action; and JEPA-based world models provide a predictive alternative for planning in latent space.

For humanoid cognitive robotics, the central opportunity is integration. A useful humanoid robot must understand instructions, perceive objects, infer hidden physical state, coordinate the whole body, and adapt under uncertainty. No single model family solves all of these requirements. The likely future is hybrid: VLA-style semantic grounding, LBM-style behavioral scale, diffusion or flow-based motor generation, and JEPA-style prediction. Such systems could move humanoid robots from impressive demonstrations toward robust, explainable and socially useful behavior.

Thus, LBMs are a promising approach for safe and reliable robotics, but not a sufficient one. They become credible safety-relevant components only when coupled with explicit constraints, uncertainty-aware monitoring, recovery policies, human override, multimodal consistency checks and evaluation protocols that measure physical risk as carefully as task success. The field should therefore move from asking whether large behavioral models are impressive to asking under which system-level safeguards they are dependable.

Acknowledgment: Supported by EU NextGenerationEU through Recovery and Resilience Plan for Slovakia, project HUROSI, number 09I01-03-V04-00048, and by Slovak Research and Development Agency, project APVV-21-0105.

References

Alansari, A. and Luqman, H. (2026). Large language models hallucination: A comprehensive survey. arXiv:2510.06265.

Asadi, M., O'Sullivan, J. W., Cao, F., Nedaee, T., Rajabalifardi, K., Li, F.-F., Adeli, E., and Ashley, E. (2026). MIRAGE: The illusion of visual understanding. arXiv:2603.21687.

Assran, M., , et al. (2025). V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv:2506.09985.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In *Conference on Computer Vision and Pattern Recognition*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Billard, A. et al. (2025). A roadmap for AI in robotics. *Nature Machine Intelligence*, 7:818–824.

Black, K. et al. (2024). π_0: A vision-language-action flow model for general robot control. arXiv:2410.24164.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. (2022). RT-1: Robotics transformer for real-world control at scale. arXiv:2212.06817.

Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., and Tenenbaum, J. B. (2026). Sycophantic chatbots cause delusional spiraling, even in ideal bayesians. arXiv:2602.19141.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, pages 15084–15097.

Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., and Song, S. (2025). Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*.

Din, M. U., Akram, W., Rosell, L. S. S. J., and Hussain, I. (2025). Vision language action models in robotic manipulation: A systematic review. arXiv:2507.10672.

Driess, D. et al. (2023). PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488.

Farkaš, I., Vavrečka, M., and Wermter, S. (2025). Will multimodal large language models ever achieve deep understanding of the world? *Frontiers in Systems Neuroscience*, 19. doi:10.3389/fnsys.2025.1683133.

Gao, C., Chen, H., Xiao, C., Chen, Z., Liu, Z., and Sun, M. (2025). H-neurons: On the existence, impact, and origin of hallucination-associated neurons in LLMs. arXiv:2512.01797.

Huang, L. et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2). doi:10.1145/3703155.

Huang, S. et al. (2023). Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems*.

Janner, M., Li, Q., and Levine, S. (2021). Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, pages 1273–1286.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12). doi:10.1145/3571730.

Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., and Fan, L. (2022). VIMA: General robot manipulation with multimodal prompts. arXiv:2210.03094.

Kalai, A. T. and Vempala, S. S. (2024). Calibrated language models must hallucinate. In *Annual ACM Symposium on Theory of Computing*, page 160–171. doi:10.1145/3618260.3649777.

Kawaharazuka, K., Oh, J., Yamada, J., and Posner, I. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. doi:10.1109/ACCESS.2025.3609980.

Khazatsky, A. et al. (2024). DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*.

Kim, M. J. et al. (2025). OpenVLA: An open-source vision-language-action model. In *8th Conference on Robot Learning*, pages 2679–2713.

LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview. https://openreview.net/forum?id=BZ5a1r-kVsf.

Li, H. et al. (2023). See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Proceedings of Machine Learning Research*, pages 1368–1378.

Li, X. et al. (2026). What matters in building vision–language–action models for generalist robots. *Nature Machine Intelligence*, 8:158–172. doi:10.1038/s42256-025-01168-7.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. In *International Conference on Learning Representations*.

Liu, S. et al. (2024). RDT-1B: A diffusion foundation model for bimanual manipulation. arXiv:2410.07864.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696.

Noda, K., Arie, H., Suga, Y., and Ogata, T. (2014). Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems*, 62(6):721–736.

NVIDIA, Bjorck, J., and Zhu, Y. (2025). GR00T N1: An open foundation model for generalist humanoid robots. arXiv:2503.14734.

Octo Model Team et al. (2024). Octo: An open-source generalist robot policy. arXiv:2405.12213.

O'Neill, A. et al. (2023). Open X-embodiment: Robotic learning datasets and RT-X models. arXiv:2310.08864.

Oquab, M. et al. (2023). DINOv2: Learning robust visual features without supervision. arXiv:2304.07193.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of Machine Learning Research*, volume 139, pages 8748–8763.

Sferrazza, C., Seo, Y., Liu, H., Lee, Y., and Abbeel, P. (2024). The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning. In *International Conference on Intelligent Robots and Systems*, pages 9698–9705.

Shridhar, M., Manuelli, L., and Fox, D. (2022). Cliport: What and where pathways for robotic manipulation. In *Proceedings of Machine Learning Research*, volume 164, pages 894–906.

Sun, Y., Sheng, D., Zhou, Z., and Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*. doi:10.1057/s41599-024-03811-x.

Terver, B., Yang, T.-Y., Ponce, J., Bardes, A., and LeCun, Y. (2025). What drives success in physical planning with joint-embedding predictive world models? arXiv:2512.24497.

TRI LBM Team et al. (2026). A careful examination of large behavior models for multitask dexterous manipulation. *Science Robotics*. doi:10.1126/scirobotics.aea6201.

Yu, X. et al. (2026). HY-embodied-0.5: Embodied foundation models for real-world agents. arXiv:2604.07430.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pages 11941–11952.

Zhao, Z. et al. (2025). Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. arXiv:2510.08807.

Zitkovich, B. et al. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of Machine Learning Research*, volume 229, pages 2165–2183.

From Remote Control to Vision-Language-Action: Initial Deployment of the Humanoid Robot Unitree G1 with Tactile Hands

Branislav Zigo

Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia
branislav.zigo@fmph.uniba.sk

Abstract

This paper reports initial deployment experience with two Unitree G1 EDU U6 humanoid robots equipped with Inspire RH56DFTP tactile five-finger hands in a university research-and-education lab. We document the path from unboxing to repeatable research use: commissioning, remote control and teleoperation, locomotion, SDK/ROS 2 configuration, integration of third-party dexterous hands, simulation-based de-risking, and cybersecurity hardening. Despite rapidly growing research demonstrations, detailed operator-facing notes on real G1 manipulation deployments remain limited; we aim to narrow this gap with a practical, system-level report. From a hardware perspective the platform largely fulfils expectations and flat-ground locomotion is convincing; nevertheless, the complete system is still fragile for everyday experimental work, especially once the dexterous hands are installed. The main friction points are software version drift and weak configuration management across vendor components, mechanical vulnerability of the external hand cabling during loco-manipulation, and the need to build a unified sensorimotor pipeline that incorporates tactile perception beyond vendor-provided hand motion primitives. We also discuss practical constraints of the onboard sensing stack (camera viewpoint and the current maturity of SLAM/navigation tooling) and summarize defensive measures required when operating a network-connected humanoid on institutional infrastructure. The paper concludes with lessons learned and a roadmap toward early Vision-Language-Action experiments with UniFoLM and GR00T.

1 Introduction

The recent availability of comparatively accessible humanoid platforms has shifted a recurring question in embodied AI from ”when will labs get a humanoid?” to ”what happens after the robot arrives?” Official material around the Unitree G1 emphasizes agility, optional secondary development, force-controlled dexterous hands, and the broader UniFoLM ecosystem (Unitree Robotics, 2024e,b; Unitree Robotics Store, 2024). At the same time, the research community is rapidly moving toward integrated Vision-Language-Action (VLA) and humanoid foundation models, including vendor-specific efforts such as Unitree UniFoLM-VLA-0 and more general systems such as NVIDIA GR00T N1 (Unitree Robotics, 2026; Bjorck et al., 2025; NVIDIA Developer, 2024). Recent G1 results already span perceptive locomotion, reduced-order MPC (Model Predictive Control) walking, single-demonstration loco-manipulation, unified whole-body control, visual sim-to-real transfer, and teleoperated medical manipulation (Sun et al., 2025; Ghansah et al., 2025; Fu et al., 2026; Sun et al., 2026; He et al., 2025; Atar et al., 2025). This makes the G1 especially interesting as a bridge platform: it is accessible enough to be seriously considered by university laboratories, yet advanced enough to expose the real systems-integration burden behind contemporary humanoid-robotics research.

This paper is intentionally written as an early deployment note. It does not claim state-of-the-art task performance; instead, it documents first practical experience from real-life testing of two Unitree G1 EDU U6 robots equipped with Inspire RH56DFTP tactile hands in a research-and-education setting. Our goal is to summarize the transition from remote control toward software-defined autonomy and, ultimately, VLA-style control. The main contribution is therefore not a new control algorithm, but an honest account of where the platform already works well, where it remains brittle, and which engineering gaps appear when one moves from promotional videos to repeatable laboratory use. While the academic literature is rapidly growing, detailed and reproducible platform-engineering reports for real manipulation deployments (especially with dexterous hands) are still relatively scarce. In contrast to many published G1 demonstrations, which rely on custom research stacks, we focus deliberately on developer workflow, configuration management, and the promise-versus-reality gap encountered during first deployment (Sun et al., 2025; Ghansah et al., 2025; Fu et al., 2026; Sun et al., 2026; He et al., 2025; Atar et al., 2025).

2 Platform Overview and Official Parameters

Officially, the G1 EDU has standing dimensions of 1320 $\times$ 450 $\times$ 200 mm, a folded size of 690 $\times$ 450 $\times$ 300 mm, a weight of about 35 kg+, and a total degree-of-freedom range of 23–43 depending on configuration (Unitree

Robotics, 2024e). The official parameter list also states up to 120 N·m knee torque, about 3 kg arm load, depth camera plus 3D LiDAR sensing, Wi-Fi 6, Bluetooth 5.2, about 2 h endurance, a 5 W speaker, and support for secondary development in the EDU edition (Unitree Robotics, 2024e), (Unitree Robotics Store, 2024). Unitree's store documentation further makes explicit that secondary development is reserved for the EDU edition rather than the base consumer configuration (Unitree Robotics Store, 2024).

For the hands, the RH56DFTP user manual specifies 12 joints, 6 DoF, 6 force sensors with 0.1 N resolution, 5–17 tactile sensors distributed across fingertips, finger pads, and the palm, repeated fingertip positioning accuracy of ±0.2 mm, and five-finger grip force of 30 N (Beijing Inspire-Robots Technology Co., Ltd., 2025). The hand operates at 24 V and exposes Ethernet, RS485, and CAN2.0 interfaces (Beijing Inspire-Robots Technology Co., Ltd., 2025). In our case, this creates a composite research platform: a Unitree humanoid body combined with third-party tactile hands, rather than a single vertically integrated end-to-end system.

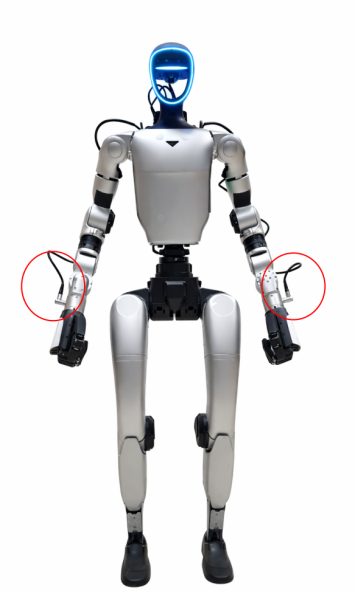


Fig. 1: Humanoid Unitree G1 EDU U6 with dexterous tactile hands RH56DFTP. Problematic external cable connection to hand containing power supply and control wires is marked by red ellipses.

3 Early Deployment Experience

3.1 Commissioning, Remote Control, and First Motion

In our early use, unpacking and basic installation were not the main problem. The robot can be brought to first motion relatively quickly if one stays close to the vendor procedure and treats the first sessions as acceptance testing rather than free exploration. The dedicated handheld controller was the most dependable interface for first commissioning. The mobile app is useful for calibration, status inspection, and service access (Unitree Robotics, 2024d), but pairing was not consistently frictionless in our tests and should be treated as an auxiliary interface rather than the primary operational one. A practical lesson is that the shortest path to safe initial use is controller first, app second, SDK third. In our case, both units were delivered within about two months from order, which is positive for academic project planning.

A second lesson concerns stair locomotion. Official demonstrations understandably highlight challenging motions, but stand-alone stair traversal should not be used as an early acceptance test in a normal laboratory environment. In our view, it remains a high-risk activity during initial deployment and should be attempted only under controlled conditions with close supervision. By contrast, ordinary flat-ground walking and running were immediately among the strongest parts of the platform. This local impression is consistent with the broader G1 literature, which already reports robust terrain locomotion and strong hardware walking performance under specialized controllers (Sun et al., 2025; Ghansah et al., 2025). In our experience, stair traversal remained unreliable and is generally not recommended during early deployment.

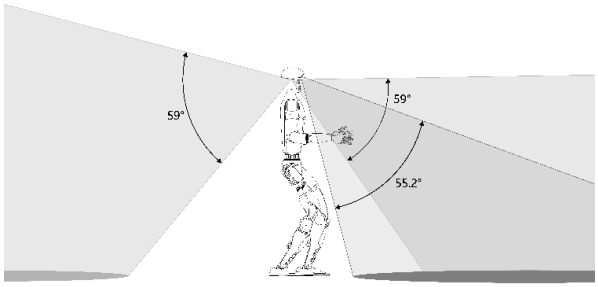


Fig. 2: Vertical range of view of RGB-D camera is limited and robot cannot see faces of humans in front of him. Source: Unitree Robotics (2024a)

3.2 Locomotion is Stronger than Manipulation

Walking and running were the most convincing robot skills in our first tests. This should, however, be interpreted carefully when manipulation is discussed. A large fraction of online promotional media features terminal covers or passive, rubber-like prosthetic hands that tolerate incidental floor contact and reduce distal mass. That setup is not equivalent to a G1 equipped with research-grade tactile five-finger hands. RH56DFTP hands are heavier, significantly more expensive, and more fragile despite their robust appearance. In addition, the external power/data cable routed from the neck switch to the hand connector is mechanically fragile and can be stressed dur-

ing manipulation and locomotion; reliable strain relief and careful cable routing are mandatory (Fig. 1). Once installed, they immediately change the practical balance envelope, acceptable collisions, floor-contact assumptions, and repair anxiety of the whole system. In short, locomotion marketing should not be naively extrapolated to locomanipulation research.

This distinction also clarifies why the same robot can look simultaneously impressive and limited. As a locomotion platform, G1 already feels mature enough for meaningful experimentation. As a manipulation platform with tactile five-finger hands, it is promising but still clearly in the systems-integration phase. The published literature supports this mixed picture: impressive real-robot loco-manipulation, visual sim-to-real, and teleoperation results do exist, but they typically rely on custom whole-body controllers, retargeting pipelines, specialized teleoperation setups, or highly tuned learning stacks rather than a stock out-of-the-box workflow (Fu et al., 2026; Sun et al., 2026; He et al., 2025; Atar et al., 2025). Cameras were among the less problematic subsystems in our early experiments, whereas the speaker was notably weaker than one would prefer for demonstrations in a noisier environment and required additional software attention.

3.3 Software Stack and Configuration Management

Here the platform reveals its largest weakness: software configuration management and version control across the vendor stack. The official development path combines SDK2, SDK2 Python, ROS 2 Foxy, CycloneDDS, and multiple Ubuntu/ROS environments (Unitree Robotics, 2024h,i,g). In practice this is not a plug-and-play experience. Parts of the stack must be built from source, environment variables such as `CYCLONEDDS_HOME` and `CMAKE_PREFIX_PATH` must be aligned, and behavior can change depending on controller/app modes and firmware versions (e.g., entering debug mode can disable higher-level motion services) (Unitree Robotics, 2025b,a). A recurring problem is documentation and sample code that refer to different (sometimes implicit) versions of firmware, middleware, and libraries, which creates unnecessary churn, hard-to-reproduce bugs, and slows down development.

From a research perspective the platform is therefore only partially open. In our deployment, the auxiliary PC2 side provided an accessible Linux development target with Jetson Orin-class compute (NVIDIA, 2024b), while the main vendor-managed PC1 side remained closed. This hybrid openness is useful but incomplete: the robot is not a single, clean, fully transparent software artifact. Without a robust producer-side versioning and release process (source, binaries, firmware, and documentation), the integration burden tends to increase as new hardware variants and software branches accumulate. In our experience, even minor upstream changes can break portability of developed software, datasets, and trained policies between otherwise similar robots, turning configuration reproducibility into a first-order engineering task.

3.4 Power Management and Development Workflow

The supplied Li-ion batteries provide roughly two hours of operation per pack in our typical lab sessions, and we observed fast charging on the order of tens of minutes (approximately 25 minutes with the supplied charger). We procured two batteries per robot and recommend this as a baseline for uninterrupted development and demonstrations. Importantly, the robot consumes nontrivial power even while standing, therefore hanging the robot on its rack during software development and testing is strongly recommended not only for safety, but also for maximizing effective uptime.

3.5 Sensors, Perception, and Navigation Tooling

The LiDAR coverage is satisfactory for basic obstacle perception, but the default RGB-D camera viewpoint is constrained by the robot's head geometry and mounting angle. As illustrated in Fig. 2, the camera primarily observes the manipulation workspace and the ground in front of the robot and cannot reliably capture faces of standing humans at conversational distance. For human-robot interaction or safety monitoring, mounting an additional camera is therefore strongly recommended. In our units, free USB-C ports are available for such extensions. Finally, the SLAM/navigation components provided in the SDK are currently rudimentary for real locomotion tasks; more mature mapping and navigation solutions should be integrated for repeatable autonomous operation in non-trivial environments.

3.6 Hand Subsystem Integration in Practice

The RH56DFTP hands (Fig. 3) are supported for basic motion generation through the vendor stack, but tactile perception is not integrated into the whole-body control pipeline out of the box. Closing the loop from tactile/force sensing to grasp stabilization and bimanual coordination therefore requires additional middleware and application-layer development.



Fig. 3: Dexterous hands RH56DFTP

4 From Remote Control to VLA

RH56DFTP is attractive precisely because it combines actuation, force sensing, and tactile sensing in one commercial hand (Beijing Inspire-Robots Technology Co., Ltd., 2025). Yet on G1 this does not automatically translate into a unified sensorimotor API. In our workflow, the body and the hands had to be treated as partially separate subsystems. This matters for contact-rich tasks: grasp closure, grasp force, tactile-state acquisition, and arm-posture synchronization do not automatically arrive as one coherent research stack. As a result, early locomanipulation is less limited by kinematics than by interface fragmentation, middleware engineering, and safety constraints. Published G1 manipulation studies point in the same direction: strong performance is possible, but it is typically achieved with custom teleoperation interfaces, grasp parameterizations, or whole-body control stacks rather than with a single unified vendor API (Fu et al., 2026; Sun et al., 2026; He et al., 2025; Atar et al., 2025).

This is also the point where tactile sensing becomes scientifically interesting. The hands promise richer feedback for grasp stability, slip, and contact events, but the practical pipeline from raw tactile/force data to a closed-loop whole-body behavior remains to be built. In our view, the body and the hands are not yet a seamless single developer platform; they are a promising multi-vendor assembly that requires explicit fusion work. That is precisely why they are interesting for academic research.

4.1 Simulation with Isaac Sim and Isaac Lab

To reduce risk, we used Isaac Sim and Isaac Lab as the main simulation route. NVIDIA describes Isaac Sim as a USD-centered (USD – Universal Scene Description), GPU-accelerated robotics simulation platform for designing, tuning, training, and deploying control agents (NVIDIA, 2024c). Isaac Lab builds on top of Isaac Sim as an open and modular framework for reinforcement learning, learning from demonstrations, and motion planning (Isaac Lab Documentation, 2024), see Fig. 4. Current NVIDIA guidance lists at least 32 GB RAM and an RTX 3070-class GPU for Isaac Sim, with higher RAM and VRAM recommended for heavier workloads and Isaac Lab training (NVIDIA, 2024a). This matches our practical impression: simulation is viable and valuable, but it is not lightweight. The main benefit is not visual polish; it is the ability to de-risk locomotion, arm motion, and data collection before touching fragile hardware.

4.2 VLA Integration Paths: UniFoLM, GR00T, and External Tooling

From the perspective of embodied AI, the most interesting shift is from remote control to VLA. Unitree now publicly exposes an embodied-AI track through its official open-source portal (Unitree Robotics, 2024b).

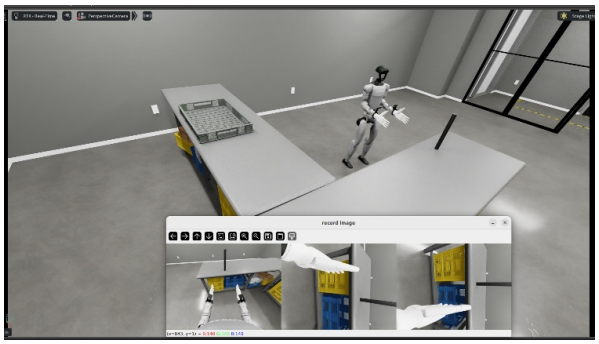


Fig. 4: Unitree G1 in Isaac Lab. Source: Unitree Robotics (2024j)

Its current UniFoLM track includes UniFoLM-VLA-0, released with training code, inference code, checkpoints, and associated G1 manipulation datasets (Unitree Robotics, 2026, 2024c). Unitree describes the model as a humanoid-manipulation VLA that extends a vision-language backbone through continued pre-training on robot-manipulation data, aiming to turn a VLM into an "embodied brain" with stronger spatial semantics and physical common sense (Unitree Robotics, 2024b, 2026). The public repository further states that the model can complete 12 categories of complex manipulation tasks in real-robot validation using a single policy (Unitree Robotics, 2026). For our purposes, the key point is not only the headline claim, but the existence of a vendor-supported path from language-conditioned perception to executable humanoid manipulation policy.

NVIDIA's GR00T should be interpreted differently. GR00T N1 is presented as an open cross-embodiment humanoid foundation model rather than a Unitree-specific stack (Bjorck et al., 2025; NVIDIA, 2025; NVIDIA Developer, 2024). Its published architecture explicitly separates a slower vision-language reasoning module (System 2) from a faster action-generation module (System 1) based on a diffusion transformer (Bjorck et al., 2025). NVIDIA further positions Isaac GR00T as a broader development platform that combines foundation models, simulation, and synthetic-data pipelines for humanoid research (NVIDIA Developer, 2024; Smith and Du, 2025). This is attractive for humanoid research because it frames adaptation not as training from scratch, but as post-training a broad prior with embodiment- and task-specific data.

Hugging Face LeRobot occupies a third role. It is neither a vendor stack nor a single VLA model, but a practical open-source layer that standardizes datasets, policies, simulation, and hardware interfaces (Hugging Face LeRobot Documentation, 2024c,b,a). Current documentation explicitly supports Unitree G1, including teleoperation, simulation, locomanipulation, and policy training, while exposing a common framework for community policies and deployment workflows (Hugging Face LeRobot Documentation, 2024c,b). The LeRobot-

Dataset v3.0 specification is especially relevant because it unifies multimodal time-series data, multi-camera video, and rich metadata in a format that can be streamed directly from the Hugging Face Hub (Hugging Face LeRobot Documentation, 2024a). In practical terms, LeRobot looks less like a final scientific claim and more like an integration substrate on which a G1 plus RH56DFTP research workflow can realistically be built.

For G1 EDU U6 equipped with RH56DFTP, however, none of these paths is plug-and-play. The core obstacle is embodiment mismatch. Public G1 examples often assume default hands, simplified end-effectors, or a different action space than a third-party tactile five-finger hand. Any serious VLA deployment will therefore require explicit redesign of observation vectors, finger-level action parameterization, control frequency, middleware bridges, and synchronized logging. In that sense, the path from remote control to VLA is real, but it passes through systems integration rather than model download alone.

4.3 Candidate Datasets for Unitree Dexterous Manipulation Training

Usable training data for Unitree dexterous manipulation can be divided into three tiers. The first tier is embodiment-near data. Unitree now publicizes G1 dexterity-hand manipulation datasets, G1 gripper manipulation datasets, Z1 dual-arm datasets, and an imitation-learning framework based on LeRobot (Unitree Robotics, 2024b,f). In addition, its current UniFoLM whole-body teleoperation collection already includes G1 whole-body tasks with Inspire hands, such as collecting clothes, picking up a pillow, and placing clothes into a washing machine or basket (Unitree Robotics, 2024c). These datasets are the most directly relevant starting point because robot morphology, control interfaces, and task kinematics are close to the target platform.

The second tier is cross-embodiment robot-learning data useful for pre-training and transfer, even if it is not directly executable on G1 plus RH56DFTP. Open X-Embodiment aggregates data from 22 robots and 527 skills (Open X-Embodiment Collaboration, 2025). DROID adds a large in-the-wild corpus of 76k manipulation trajectories collected across 564 scenes (Khazatsky et al., 2024). RH20T is especially interesting for contact-rich manipulation because it provides more than 110k real-world robot sequences with visual, force, audio, and action signals, together with paired human demonstrations and language descriptions (Fang et al., 2024). For a humanoid dexterous-hand project, such datasets are best viewed as sources of visual-language-manipulation priors, subtask structure, and robustness, not as drop-in finger-control supervision.

The third tier comprises hand-object interaction datasets that are most useful for perception, retargeting, grasp priors, and contact modeling. DexYCB focuses on hand grasping of objects and benchmarks 6D object pose, 3D hand pose, and handover-relevant grasp generation (Chao et al., 2021). HOI4D provides 2.4 million RGB-D egocentric frames with hand pose, object pose, action, and scene annotations (Liu et al., 2022). ARCTIC concentrates on dexterous bimanual hand-object manipulation with articulated objects and dynamic contact over 2.1 million frames (Fan et al., 2023). GRAB adds whole-body grasping sequences with contact information, hand pose, object pose, and full-body motion (Taheri et al., 2020). These are not end-to-end Unitree policy datasets, but they are highly relevant for building the perception and retargeting layers that dexterous humanoid manipulation still lacks.

Consequently, the most credible data strategy for our platform is layered: start from vendor- and embodiment-near Unitree datasets, add laboratory teleoperation data collected on the actual RH56DFTP-equipped robots, and then use large cross-embodiment corpora plus hand-object datasets for pre-training, perception, and synthetic-data augmentation. This hybrid strategy is more realistic than expecting one open dataset to solve whole-body humanoid dexterous manipulation end-to-end.

5 Cybersecurity and Operational Deployment Implications

Humanoid deployment is not only a robotics problem but also a networked-systems problem. In a university environment, the robot is routinely connected to developer laptops, institutional networks, external cameras, and storage. This enlarges the attack surface and turns the humanoid into a high-value cyber-physical endpoint that can influence both information security and physical safety. A recent SoK (Systematization of Knowledge) on humanoid cybersecurity evaluates G1 EDU among real platforms and argues that humanoids require a layered security model spanning hardware, operating system, middleware, cloud services, and human-robot interaction (Surve et al., 2025). In addition, a dedicated G1 security assessment highlights weaknesses in provisioning, cryptographic material reuse, and telemetry behavior that motivate a conservative operational stance for laboratory deployments (Mayoral-Vilches et al., 2025).

5.1 Reported Attack Surfaces and Data Flows

The G1 exposes multiple concurrent communication surfaces: local Ethernet/Wi-Fi networking used for SDK and ROS 2 operation, Bluetooth Low Energy (BLE) used for provisioning and mobile-app interaction, and additional runtime services for media streaming and cloud connectivity. On the local network, DDS/RTPS traffic enables high-rate publish/subscribe messaging between components and external development machines; with-

out additional protections, this traffic can be observed or interfered with by any device on the same segment. In parallel, cloud-facing channels may be used for telemetry, over-the-air updates, and remote services. For laboratories, the practical implication is that physical isolation alone is insufficient: the robot must be treated like a managed endpoint with explicit network policy and monitoring.

5.2 Key Issues Highlighted in G1 Security Analyses

Mayoral-Vilches et al. (2025) report that the BLE-based provisioning workflow can be abused to obtain privileged code execution when an attacker is in Bluetooth range, and that the corresponding cryptographic material is reused across devices, reducing effective security. The same assessment documents multiple active communication channels (e.g., local DDS/RTPS buses, cloud telemetry/OTA coordination, and media streaming services) and reports weaknesses in the proprietary configuration protection scheme (Mayoral-Vilches et al., 2025). In particular, it describes periodic transmission of detailed robot state and system telemetry to external servers and notes implementation choices that may weaken transport-layer protections in some clients (Mayoral-Vilches et al., 2025). While the exact impact depends on firmware version and deployment configuration, these results justify a conservative operational stance: assume that (i) provisioning interfaces can be a compromise vector, (ii) default network services can leak sensitive sensor and state information unless explicitly constrained, and (iii) security posture can change across software/firmware updates and should be revalidated after each upgrade.

5.3 Practical Mitigations for Academic Deployments

Based on our operational needs and the above threat model, we recommend a defensive baseline for laboratory use: (1) place the robot on a dedicated VLAN or behind a small router with explicit egress filtering; (2) disable Bluetooth and unnecessary services when not required for experiments; (3) implement a default-deny outbound firewall policy and allow only explicitly needed endpoints (e.g., time synchronization, approved model APIs) while blocking telemetry/OTA paths when data sovereignty is required; (4) monitor traffic (packet captures or flow logs) during initial commissioning and after any firmware update; and (5) keep a documented update governance process (who updates, when, and how changes are validated) to avoid accidental regression of both safety and security settings. These practices are particularly important for VLA experiments, where high-level autonomy must be paired with strict safety interlocks and strong network hygiene.

For our deployments, the key lesson is that security controls must be designed into the daily workflow: networking, logging, and access control are as important as kinematics and perception when the robot is connected to institutional infrastructure. Treating cybersecurity as part of the acceptance test (together with locomotion and manipulation) reduces the risk of silent data leakage and improves reproducibility of subsequent experiments.

6 Conclusion

These first experiments suggest that Unitree G1 EDU with RH56DFTP hands is not a turnkey humanoid worker, but it is already a valuable research platform. Its strongest asset is not that every subsystem is mature; it is that the platform makes current embodied-AI questions concrete and experimentally tractable. Walking and running are already convincing, and the broader literature confirms that the platform can support considerably more advanced locomotion, loco-manipulation, and teleoperation under custom research stacks (Sun et al., 2025; Ghansah et al., 2025; Fu et al., 2026; Sun et al., 2026; He et al., 2025; Atar et al., 2025). The main research friction lies elsewhere: software integration, middleware reproducibility, safe use of fragile dexterous hands, tactile-data fusion, and the passage from remote control to unified autonomous control.

In that sense the platform is best understood as a transition technology. It brings humanoid research out of the purely aspirational phase and into everyday laboratory engineering. We therefore view it as one of the first comparatively accessible commercially available humanoid setups that can realistically enter university research and education, while also making it impossible to ignore how much work remains between impressive demos and dependable autonomous locomanipulation.

7 Future Work

Our near-term work focuses on making the platform more reproducible and less brittle for manipulation experiments. This includes hardening the software stack with explicit version pinning and configuration snapshots, improving mechanical strain relief for the external hand cabling, and extending the sensing setup with additional cameras for human-facing perception. On the autonomy side, we plan controlled experiments with UniFoLM-VLA and NVIDIA GR00T, starting with offline evaluation and teleoperation dataset collection, followed by safety-constrained on-robot rollouts. A key technical goal is to integrate RH56DFTP tactile and force signals into the observation and control loop (e.g., slip and contact-event feedback for grasp stabilization), and to replace the rudimentary SDK navigation with a more capable SLAM and locomotion planning pipeline suitable for repeatable locomanipulation in real environments.

Acknowledgement

This project has been funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00048.

References

Atar, S., Liang, X., Joyce, C., Richter, F., Ricardo, W., Goldberg, C., Suresh, P., and Yip, M. (2025). Humanoids in hospitals: A technical study of humanoid robot surrogates for dexterous medical interventions.

Beijing Inspire-Robots Technology Co., Ltd. (2025). The dexterous hand RH56DFTP user manual, V1.0.0. Available online: https://en.inspire-robots.com/wp-content/uploads/2025/01/INSPIRE-ROBOTS-The-Dexterous-Hand-RH56DFTP-User-Manual-V1.0.0.pdf.

Bjorck, J. et al. (2025). GR00T N1: An open foundation model for generalist humanoid robots.

Chao, Y.-W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y. S., Van Wyk, K., Iqbal, U., Birchfield, S., Kautz, J., and Fox, D. (2021). DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049.

Fan, Z. et al. (2023). ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*. doi: 10.1109/CVPR52729.2023.01244.

Fang, H.-S., Fang, H., Tang, Z., Liu, J., Wang, C., Wang, J., Zhu, H., and Lu, C. (2024). RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660.

Fu, Y., Xie, F., Xu, C., Xiong, J., Yuan, H., and Lu, Z. (2026). DemoHLM: From one demonstration to generalizable humanoid loco-manipulation. *IEEE Robotics and Automation Letters*, 11(4):4393–4400.

Ghansah, A. B., Esteban, S. A., and Ames, A. D. (2025). Hierarchical reduced-order model predictive control for robust locomotion on humanoid robots. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pages 1–8.

He, T., Wang, Z., Xue, H., Ben, Q., Luo, Z., Xiao, W., Yuan, Y., Da, X., Castañeda, F., Sastry, S., Liu, C., Shi, G., Fan, L., and Zhu, Y. (2025). VIRAL: Visual sim-to-real at scale for humanoid loco-manipulation.

Hugging Face LeRobot Documentation (2024a). LeRobotDataset v3.0. Available online: https://huggingface.co/docs/lerobot/lerobot-dataset-v3.

Hugging Face LeRobot Documentation (2024b). Policy deployment (lerobot-rollout). Available online: https://huggingface.co/docs/lerobot/main/inference.

Hugging Face LeRobot Documentation (2024c). Unitree G1. Available online: https://huggingface.co/docs/lerobot/main/en/unitree_g1.

Isaac Lab Documentation (2024). Overview. Available online: https://isaac-sim.github.io/IsaacLab/v1.4.1/.

Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Sridhar, A., Zhu, Y., Geng, Q., Han, J., Chen, D., et al. (2024). DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*.

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., and Yi, L. (2022). HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022.

Mayoral-Vilches, V., Makris, A., and Finisterre, K. (2025). Cybersecurity ai: Humanoid robots as attack vectors.

NVIDIA (2024a). Isaac Sim technical requirements and minimum system requirements. Available online: https://docs.omniverse.nvidia.com/isaacsim/latest/common/technical-requirements.html.

NVIDIA (2024b). Jetson Orin technical specifications. Available online: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/.

NVIDIA (2024c). What is Isaac Sim? Available online: https://docs.omniverse.nvidia.com/isaacsim/latest/overview.html.

NVIDIA (2025). NVIDIA announces Isaac GR00T N1 - the world's first open humanoid robot foundation model - and simulation frameworks to speed robot development. Press Release. Available online: https://investor.nvidia.com/news/press-release-details/2025/NVIDIA-Announces-Isaac-GR00T-N1---the-Worlds-First-Open-Humanoid-Robot-Foundation-Model--and-Simulation-Frameworks-to-Speed-Robot-Development/default.aspx.

NVIDIA Developer (2024). Isaac GR00T. Available online: https://developer.nvidia.com/isaac/gr00t.

Open X-Embodiment Collaboration (2024). Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903.

Open X-Embodiment Collaboration (2025). Open X-Embodiment: Robotic learning datasets and RT-X Models.

Smith, C. and Du, P. (2025). Building a synthetic motion generation pipeline for humanoid robot learning. NVIDIA Technical Blog. Available online: https://developer.nvidia.com/blog/building-a-synthetic-motion-generation-pipeline-for-humanoid-robot-learning/.

Sun, W., Cao, B., Chen, L., Su, Y., Liu, Y., Xie, Z., and Liu, H. (2025). Learning perceptive humanoid locomotion over challenging terrain. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6571–6578.

Sun, W., Feng, L., Cao, B., Liu, Y., Jin, Y., and Xie, Z. (2026). ULC: A unified and fine-grained controller for humanoid loco-manipulation.

Surve, P. P., Shabtai, A., and Elovici, Y. (2025). SoK: Cybersecurity assessment of humanoid ecosystem.

Taheri, O., Ghorbani, N., Black, M. J., and Tzionas, D. (2020). GRAB: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer.

Unitree Robotics (2024a). G1 SDK development guide. Available online: https://support.unitree.com/home/en/G1_developer.

Unitree Robotics (2024b). Official open source portal. Available online: https://www.unitree.com/opensource/.

Unitree Robotics (2024c). UnifoLM_WBT_Dataset. Hugging Face collection. Available online: https://huggingface.co/collections/unitreerobotics/unifolm-wbt-dataset.

Unitree Robotics (2024d). Unitree explore APP download – G1 tutorial videos and manuals. Available online: https://www.unitree.com/mobile/app/g1.

Unitree Robotics (2024e). Unitree G1. Product page. Available online: https://www.unitree.com/g1/.

Unitree Robotics (2024f). unitree_IL_lerobot. GitHub repository. Available online: https://github.com/unitreerobotics/unitree_IL_lerobot.

Unitree Robotics (2024g). unitree_ros2. GitHub repository. Available online: https://github.com/unitreerobotics/unitree_ros2.

Unitree Robotics (2024h). unitree_sdk2. GitHub repository. Available online: https://github.com/unitreerobotics/unitree_sdk2.

Unitree Robotics (2024i). unitree_sdk2_python. GitHub repository. Available online: https://github.com/unitreerobotics/unitree_sdk2_python.

Unitree Robotics (2024j). unitree_sim_isaaclab. GitHub repository. Available online: https://github.com/unitreerobotics/unitree_sim_isaaclab.

Unitree Robotics (2025a). Instructions to control G1 movement via SDK - do not enter debug mode. GitHub Issue #43. Available online: https://github.com/unitreerobotics/unitree_sdk2_python/issues/43.

Unitree Robotics (2025b). Problems with cyclonedds while using NVIDIA Jetson. GitHub Issue #53. Available online: https://github.com/unitreerobotics/unitree_sdk2_python/issues/53.

Unitree Robotics (2026). UnifoLM-VLA-0: A vision-language-action (VLA) framework under UnifoLM family. GitHub repository. Available online: https://github.com/unitreerobotics/unifolm-vla.

Unitree Robotics Store (2024). Unitree G1 store page and EDU note on secondary development. Available online: https://shop.unitree.com/en-ca/products/unitree-g1.

Index of authors

Author Index

Symbols

Čeněk, Jiří 26
Černý, David 29
Červenka, Adam 21
Ďalak, Frederik 39
Šišková, Julie 135
Šmejkal, Jan 120
Štěpánová, Karla 156

A

Al Jallad, Mhd Walid 12
Andrejkova, Gabriela 17

B

Behrens, Jan Kristof 156
Bílá, Tamara 21

C

Chromková Manea, Beatrice Elena 26
Chrásková, Noemi 135
Cibula, Miroslav 34

D

Debevec-Kruse, Aly 41
Dlouhá, Jana 78
Dlouhý, Matyáš 147
Dvořák, Michal 147
Děchtěrenko, Filip 135

E

Eklová, Laura 135
Epifanic, Jelena 46

F

Farkaš, Igor 21, 34, 52, 104, 178

G

Gregor, Radovan 52

H

Hananeia, Nicholas 56
Havlík, Kryštof 61
Hlaváč, Václav 156
Hoffmann, Matěj 69
Hromková, Lenka 26
Hübelová, Dana 26

J

Jedlička, Peter 143
Jonáš, Juraj 73

K

Kapounová, Ivana 78
Kiš, Milica 85
Kollárová, Hana 90
Kopčo, Norbert 17
Kostka, Nikita 97
Kostrhun, Michal 104
Krajčířová, Anna 108

L

Lúčny, Andrej 116

M

Macík, Miroslav 61
Malinovská, Kristína 12, 34, 46
Mokrý, Stanislav 26
Mužíková, Leona 26

S

Stanová, Magda 124
Suchodolová, Veronika 26
Svoboda, Aleš 129
Svozilová, Anna 26

T

Takáč, Martin 41, 85, 138

Tomko, Matúš 143

V

Vadinský, Ondřej 97, 147
Vanc, Petr 156

W

Wiedermannn, Jiří 163

Y

Yarzhemska, Anastasiia 167

Z

Zamrazilová, Karina 174
Zigo, Branislav 178, 187

Juraj Hvorecký
Karina Zamrazilová
Kristína Malinovská
Kryštof Havlík
editors

Kognice a umělý život
Cognition and Artificial Life
KUŽ / CAL 2026

Published by Czech Technical University in Prague, 2026

Title page design: Julia Kovalčíková and Kryštof Havlík

Extent 197 pages, first edition,
published as an electronic publication

ISBN 978-80-01-07572-2