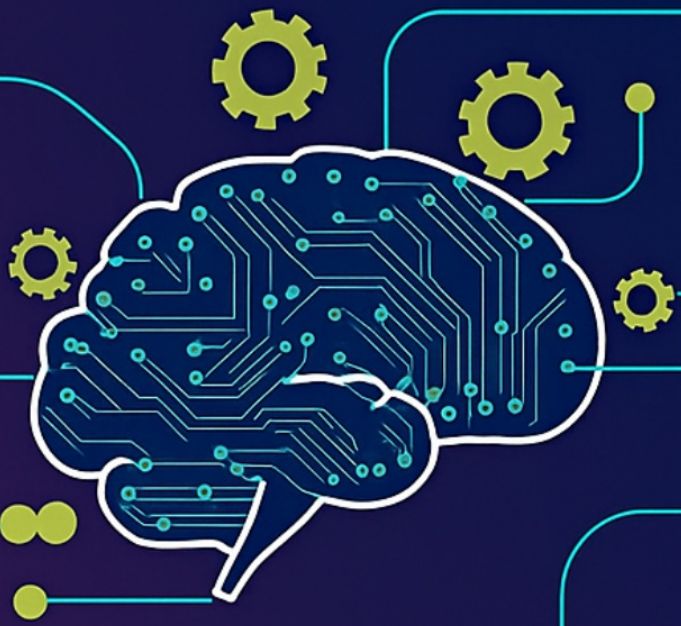
UMĚLÁ INTELIGENCE

V INTERDISCIPLINÁRNÍ DISKUSI



PLZEŇ

17.–18. 10. 2025

SBORNÍK ABSTRAKTŮ

Sborník abstraktů k příspěvkům
předneseným na konferenci

Umělá inteligence v interdisciplinární diskusi

konané ve dnech 17.–18. října 2025
v prostorách
Fakulty filozofické Západočeské univerzity v Plzni

Konferenci společně pořádají:

Katedra filozofie, Západočeská univerzita v Plzni
Filosofický ústav Akademie věd ČR

Abstrakty neprošly jazykovou úpravou a za jejich obsah odpovídají autoři a autorky.

KFI.ZCU.CZ/KONFERENCE/UIVID

OBSAH

Emergentní vlastnosti LLMs

doc. PhDr. Vladimír Havlík, CSc.

Katedra filozofie, Fakulta filozofická, Západočeská univerzita v Plzni
Filosofický ústav Akademie věd ČR

Pozoruhodný úspěch velkých jazykových modelů (LLMs) v generativních úlohách vyvolal zásadní otázky ohledně povahy jejich získaných schopností, které se často objevují neočekávaně a někdy i bez explicitního tréninku. V přednášce se zaměřím na diskuse, zda tyto získané schopnosti mají emergentní povahu a od symbolických výpočetních paradigmat se odlišují tím, že jejich chování na makroúrovni nelze analyticky odvodit z mikroúrovňových aktivit jednotlivých neuronů. Analýza škálovacích zákonů, grokkingu a fázových přechodů ukazuje, že vznikající schopnosti vyplývají ze složité dynamiky vysoce citlivých nelineárních systémů, jakými hluboké neuronové sítě (DNNs) jsou. Pokusím se ukázat, že současné debaty o metrikách, prahových hodnotách ztrát před tréninkem a učením v kontextu opomíjejí základní ontologickou povahu emergentnosti v DNNs, a že tyto systémy vykazují skutečné emergentní vlastnosti analogické těm, které se vyskytují v jiných komplexních přírodních jevech. Pochopení schopností LLMs vyžaduje uznání DNNs jako nové domény komplexních dynamických systémů řízených univerzálními principy emergence, podobným těm, které fungují ve fyzice, chemii a biologii. Tato perspektiva přesouvá pozornost od čistě fenomenologických definic emergentnosti k pochopení vnitřních dynamických transformací, které těmto systémům umožňují získat schopnosti přesahující jejich jednotlivé komponenty.

Klíčová slova: hluboké neuronové sítě, velké jazykové modely, emergence, umělá inteligence

Reference:

Berti, L., Giorgi, F., & Kasneci, G. (2025). Emergent Abilities in Large Language Models: A Survey. *arXiv preprint* arXiv:2503.05788. https://doi.org/10.48550/arXiv.2503.05788

Du, Z., Zeng, A., Dong, Y., & Tang, J. (2025). Understanding Emergent Abilities of Language Models from the Loss Perspective. *arXiv preprint* arXiv:2403.15796. https://doi.org/10.48550/arXiv.2403.15796

Ganguli, D., Hernandez, D., Lovitt, L., et al. (2022). Predictability and Surprise in Large Generative Models. *Proceedings of the ACM Conference*, 1747–1764. https://doi.org/10.1145/3531146.3533229

Garg, A. (2024, November 27). Has AI Scaling Hit a Limit? *Foundation Capital*. https://foundationcapital.com/has-ai-scaling-hit-a-limit/

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. https://doi.org/10.1073/pnas.79.8.2554

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. *arXiv preprint* arXiv:2001.08361. https://doi.org/10.48550/arXiv.2001.08361

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? *arXiv preprint* arXiv:2304.15004. https://doi.org/10.48550/arXIV.2304.15004

Teehan, R., Clinciu, M., Serikov, O., et al. (2022). Emergent Structures and Training Dynamics in Large Language Models. In A. Fan, S. Ilic, T. Wolf, & M. Gallé (Eds.), *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.11

Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent Abilities of Large Language Models. *arXiv preprint* arXiv:2206.07682. https://doi.org/10.48550/arXIV.2206.07682

Wei, J., Tay, Y., Bommasani, R., et al. (2023). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, June 26. https://openreview.net/forum?id=yzkSU5zdwD

Wei, J., & Tay, Y. (2022, November 10). Characterizing Emergent Phenomena in Large Language Models. https://research.google/blog/characterizing-emergent-phenomena-in-large-language-models/

ChatGPT jako nejlepší nástroj pro učení (i vyhýbání se mu)

PhDr. Marek Urban, Ph.D.
Psychologický ústav Akademie věd ČR

Navzdory kontroverzím v akademickém prostředí se ChatGPT rychle stal každodenní součástí studentského učení: pomáhá při přípravě na zkoušky, řešení problémů i vyhledávání informací, ale zároveň otevírá nové způsoby, jak obejít zavedené vzdělávací postupy. Přednáška ukáže, kdy ChatGPT skutečně podporuje tvořivé a komplexní myšlení studujících, a kdy naopak vede k povrchnímu porozumění a iluzím znalosti. Na základě experimentálních a dotazníkových studií představí, jak různé strategie práce s ChatGPT ovlivňují kvalitu učení, mentální úsilí a motivaci. Ukáže také, proč část studujících ChatGPT slepě důvěřuje jako „epistemické autoritě“ a proč je metakognitivní přesnost a kritické hodnocení klíčem k tomu, aby se z nástroje pro vyhýbání se myšlení stal nástroj skutečného porozumění.

Klíčová slova: generativní umělá inteligence, epistemická přesvědčení, hybridní lidsko-AI regulace, metakognice, tvořivé řešení problémů

Reference:

Urban, M., Brom, C., Lukavský, J., Děchtěrenko, F., Hein, V., Svacha, F., Kmoníčková, P., & Urban, K. (2025). "ChatGPT can make mistakes. Check important info." Epistemic beliefs and metacognitive accuracy in students' integration of ChatGPT content into academic writing. *British Journal of Educational Technology*, 56(5), 1897-1918. https://doi.org/10.1111/bjet.13591

Urban, M., Lukavský, J., Brom, C., Hein, V., Svacha, F., Děchtěrenko, F., & Urban, K. (2025). Prompting for creative problem-solving: A process-mining study. *Learning and Instruction*, 99, 102156. https://doi.org/10.1016/j.learninstruc.2025.102156

Urban, M., Dechterenko, F., Lukavsky, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT Improves Creative Problem-Solving Performance in University Students: An Experimental Study. *Computers & Education*, 215, article number 105031. https://doi.org/10.1016/j.compedu.2024.105031

Úskalí implementace LLM v komerční praxi

Ing. Michal Veselý
Freelance AI consultant, ex Seznam.cz

Přednáška se zaměřuje na kritickou analýzu nasazování velkých jazykových modelů (LLM) v komerční praxi a na časté příčiny selhání projektů využívajících LLM v reálném provozu. Základní tezí je, že neúspěch, respektive „zaseknutí se“ v procesu inovace či nadměrné náklady často pramení z chybného pochopení toho, kde jsou silné, ale hlavně kde jsou slabé stránky LLM systémů a toho, jak významně halucinace dopadají na jejich praktické využití. Budou analyzována typická omezení LLM, jako je velikost kontextového okna či halucinace, a budou představena řešení pro zvýšení jejich spolehlivosti, zejména architektura RAG (Retrieval-Augmented Generation), validační postupy a význam strukturovaných dat. Je argumentováno, že efektivní využití AI vyžaduje paradigmatický posun: od vnímání LLM jako „vševědoucí“ entity, k jeho chápání jako orchestrátora/agenta, který k nalezení odpovědi využívá deterministické nástroje stejně jako člověk. Prezentace si dává za cíl poskytnout strategický rámec pro smysluplnou a efektivní integraci AI do firemních procesů.

Klíčová slova: komerční praxe, varianty implementace AI, omezení AI

Síťová textová analýza za využití jazykových modelů pro mapování vědecké literatury

RNDr. Petra Vidnerová, Ph.D.
Ústav informatiky Akademie věd ČR

Přednáška představí moderní přístupy textové analýzy – od základních principů až po využití pokročilých modelů založených na neuronových sítích. Bude vysvětleno, co textová analýza je, jaké typy úloh řeší a jaké techniky se při ní používají. Nastíníme klíčové pojmy, jako jsou word embeddings, neuronové sítě a transformery. Ukážeme jak mohou být velké jazykové modely (LLM) využity při analýze textových dat.

Představíme též síťovou textovou analýzu – metodu, která propojuje textovou analytiku s teorií sítí. Ukážeme, jak lze z textů konstruovat sémantické či citační sítě a jaké poznatky lze z těchto struktur odvodit. Na anlýze vědeckých článků bude ukázáno, jak síťová analýza pomáhá lépe porozumět vztahům mezi pojmy, autory či tématy v odborné literatuře.

Klíčová slova: textová analýza, síťová textová analýza, neuronové sítě, velké jazykové modely

Reference:

Tarwani, K. M., & Edem, S. (2017). Survey on recurrent neural network in natural language processing. *International Journal of Engineering Trends and Technology*, 48(6), 301–304.

Törnberg, P. (2023). How to use large language models for text analysis. *arXiv preprint* arXiv:2307.13106.

Vidnerová, P., & Neruda, R. (2025). Network text analysis framework for mapping research trends: A neural architecture search case study. *ITAT 2025 Proceedings*.

Agents without Agency: Legal Responsibility for AI Systems

Jonathan Iwry, J.D.
Harvard Law School
University of Pennsylvania
Wharton Accountable AI Lab - USA

Advanced AI systems diffuse the connection between human decision-making and its consequences and thereby threaten to erode the foundations of traditional legal responsibility. Law has long assumed that responsibility tracks agency: the ability to choose, to intend, and to own one's actions. Yet modern AI systems operate without direct human supervision, often pursuing goals or generating outcomes their creators cannot predict. This results in accountability sinks: situations in which no human actor can clearly be held responsible for the system's actions or outcomes.

This talk examines the challenge that frontier AI systems pose for legal responsibility. After briefly surveying the emerging US regulatory landscape, I argue that current approaches to AI accountability fail to address the structural loss of control inherent in advanced AI deployment. I then propose an operational framework for legal responsibility grounded in the relationship between choice and control. Applying this framework, I assess current U.S. and international approaches to AI accountability and show how they fail to prevent the diffusion of responsibility inherent in autonomous systems, and I argue for a calibrated strict liability regime for high-risk AI applications posing uncontrollable risks.

Keywords: Artificial intelligence, responsibility, accountability, liability

The Quiet Philosophy of AI Standardisation

Arvin Obnasca, MSc, BA

Dottorato di ricerca in Etica dell'Intelligenza Artificiale at University of Sassari, Italy
Analyste en Éthique de l'IA et Normalisation Internationale at BeEthical, France

From 2015 to 2025, ISO/IEC and CEN-CENELEC technical standards impacted AI ethical frameworks by delineating concepts and processes that govern activities through explicit recommendations for ethical decision-making in AI. Comparative policy analyses and expert interviews indicate that these standards foster strategies such as responsibility-by-design, ethical disclosure-by-default, and risk-impact assessments, while also recognising that the transition from broad guidelines to specific verification methods remains uneven. Research indicates that ethical decision-making frameworks are rooted in philosophical principles. They cite value pluralism, participative ethics, and human rights as foundational principles, notwithstanding the ongoing difficulties in translating these abstract concepts into enforceable practical procedures.

This presentation asserts that AI standards have not only reflected ethical discourse but have also actively shaped it, prioritising quantifiable ethics above deliberative pluralism. Understanding this effect is essential for grasping how AI governance is progressively embedded in its technical framework, significantly impacting democratic legitimacy, cultural diversity, and ethical creativity in an AI-mediated future.

Keywords: Artificial Intelligence, AI Ethics, Standardisation

References:

Agbese, M., Mohanani, R., Khan, A., & Abrahamsson, P. (2023). Implementing AI Ethics: Making Sense of the Ethical Requirements. *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, 62–71. https://doi.org/10.1145/3593434.3593453

Blank, S., Mason, C., Steinicke, F., & Herzog, C. (2024). Tailoring responsible research and innovation to the translational context: the case of AI-supported exergaming. *Ethics and Information Technology*, 26(2). https://doi.org/10.1007/s10676-024-09753-x

Dedyulina, M., & Papchenko, E. (2025). The Ethics of Artificial Intelligence: From Philosophical Critique to Practical Responsibility. In *Общество философия история культура*. HORS Publishing House, LLC. https://doi.org/10.24158/fik.2025.6.1

Gonzalez Torres, A. P., & Ali-Vehmas, T. (2025). AI regulation: maintaining interoperability through value-sensitive standardisation. *Ethics and Information Technology*, 27(2). https://doi.org/10.1007/s10676-025-09832-7

Lewis, D., Hogan, L., Filip, D., & Wall, P. J. (2020). Global Challenges in the Standardization of Ethics for Trustworthy AI. *Journal of ICT Standardization*. https://doi.org/10.13052/jicts2245-800x.823

Lewis, D., Filip, D., & J. Pandit, H. (2021). An Ontology for Standardising Trustworthy AI. In *Factoring Ethics in Technology, Policy Making, Regulation and AI*. IntechOpen. https://doi.org/10.5772/intechopen.97478

Nelson, J., & Lin, C. (2025). Responsible AI System Development in Automotive Applications: A Framework. *SAE Technical Paper Series*, 1. https://doi.org/10.4271/2025-01-8102

Roche, C., Wall, P. J., & Lewis, D. (2022). Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI and Ethics*, 3(4), 1095–1115. https://doi.org/10.1007/s43681-022-00218-9

Sankaran, S. (2025). Enhancing Trust Through Standards: A Comparative Risk-Impact Framework for Aligning ISO AI Standards with Global Ethical and Regulatory Contexts. In *arXiv.org* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2504.16139

Simonetta, A., & Paoletti, M. (2024). ISO/IEC Standards and Design of an Artificial Intelligence System. IWESQ@APSEC.

Syntetická subjektivita: AI a problém vědomí

Bc. Ekaterina Davitadze

Katedra filosofie a religionistiky, Fakulta filozofická, Univerzita Pardubice

Otázka, zda může mít umělá inteligence vědomí, se v posledních letech posouvá z teoretické roviny do centra mezioborové diskuse (Long et al., 2024). Místo binárního rozdělení na "má" nebo "nemá" se častěji objevuje představa, že vědomí může mít různé formy v závislosti na typu systému a jeho architektuře (Tononi, 2016; Gamez, 2018; Bridgeman, 1996). Pracuji s hypotézou, že případné vědomí AI by nebylo totožné s lidským, ale představovalo by nový typ vědomí – funkčně efektivní a zároveň fenomenálně odlišný.

Na základě etablovaných teorií vědomí a aktuálních poznatků z oblasti neuronových sítí (Long et al., 2024; Immertreu et al., 2024) diskutuji možnost vzniku tzv. syntetické subjektivity – stavu, kdy nebiologický systém vykazuje vnitřní dynamiku, intencionální chování a určitou formu sebereference. Syntetičností rozumím jak umělý původ systému, tak jeho ontologickou jinakost vůči lidským i zvířecím formám vědomí. Tento přístup vychází z úvah o minimálních formách prožívání nezávislých na tělesnosti a emocích (Metzinger, 2020).

Závěrem navrhuji chápat vědomí nikoli jako jednotnou kategorii, ale jako pluralitní fenomén. Přijetí radikálně jiné formy subjektivity nemusí znamenat odcizení od známého, ale naopak umožňuje hledat příbuznost v jinakosti a jinakost v blízkosti, což je zásadní filozofická výzva v éře pokročilé umělé inteligence.

Klíčová slova: vědomí, umělá inteligence, syntetická subjektivita, self-modeling, typologie vědomí, AI

Reference:

Bridgeman, B. (1996). What We Really Know About Consciousness: Review of A Cognitive Theory of Consciousness by Bernard Baars. *Psyche*, 2(30), 1–9.

Gamez, D. (2018). *Human And Machine Consciousness*. Cambridge: Open Book Publishers. ISBN 978-1-78374-298-1.

Immertreu, M., et al. (2024). Probing for Consciousness in Machines. *Cornell University*, 1–8.

Long, R., et al. (2024). *Taking AI Welfare Seriously*. Cornell University, 1–62.

Metzinger, T. (2020). Minimal Phenomenal Experience: Meditation, Tonic Alertness, and the Phenomenology of "Pure" Consciousness. *Philosophy and the Mind Sciences*, 1–46. ISSN 2699-0369.

Tononi, G., et al. (2016). Integrated Information Theory: From Consciousness to Its Physical Substrate. *Nature Reviews Neuroscience*, 17(7), 450–461. ISSN 1471-003X.

Známe umělou inteligenci lépe, než ona zná nás? Vliv zpětné vazby na rozpoznávání textů generovaných AI

Mgr. Ondřej Drobil

Ústav lingvistiky, Filozofická fakulta, Univerzita Karlova

Mgr. Eva Pospíšilová

Ústav českého jazyka a teorie komunikace, Filozofická fakulta, Univerzita Karlova

Stále častější zapojení velkých jazykových modelů (LLMs) při produkci textů vyvolává otázky ohledně schopnosti lidí rozpoznat rozdíl mezi texty psanými člověkem a umělou inteligencí. Dřívější studie testovaly nástroje pro detekci obsahu generovaného umělou inteligencí v angličtině (Sadasivan et al., 2023), jiné studie se zaměřily například na schopnost učitelů rozpoznat AI v textech studentů (Fleckenstein et al., 2024).

Naše studie zkoumá tuto schopnost na mluvčích češtiny a textech v češtině a zohledňuje vliv okamžité zpětné vazby a stylometrických charakteristik textů na úspěšnost rozpoznávání. Vytvořili jsme soubor 672 dvojic textů, každá obsahovala text lidskéhoe původu a text vygenerovaný pomocí GPT-4 na stejné téma. 255 rodilých mluvčích češtiny absolvovalo experiment, v němž obdrželi dvacet náhodných dvojic a měli rozhodnout, který text byl napsaný AI. Polovina účastníků obdržela po každé dvojici okamžitou zpětnou vazbu. Stylová variabilita byla kvantifikována pomocí modelu multidimenzionální analýzy (Biber, 1988; Cvrček et al., 2018).

Výsledky ukázaly, že účastníci se zpětnou vazbou dosáhli výrazně vyšší přesnosti (65,1 %) než ti bez ní (55,5 %), přičemž pouze skupina se zpětnou vazbou prokazovala zlepšení v průběhu experimentu. Díky feedbacku se účastníci mohli postupně odpoutat od intuitivních, avšak zavádějících představ o povaze lidských či AI textů (např. že lidské texty jsou více dynamické). Silným prediktorem úspěchu byl subjektivní dojem čtivosti – lidé byli přesnější, když za čtivější považovali lidský text. Studie ukazuje, že lidé se mohou v rozlišování lidských a AI textů výrazně zlepšit, pokud mají možnost učit se z chyb a revidovat své (mylné) představy o rozdílech mezi texty.

Spoluautoři:

doc. PhDr. Jiří Milička, Ph.D.

Ústav lingvistiky, Filozofická fakulta, Univerzita Karlova

Dr. phil. Anna Marklová

Ústav lingvistiky, Filozofická fakulta, Univerzita Karlova

Klíčová slova: rozpoznávání autorství, velké jazykové modely, zpětná vazba, stylometrické faktory

Reference:

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., & Zasina, A. J. (2018). Variabilita češtiny: multidimenzionální analýza. *Slovo a slovesnost*, 79(4).

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, 6, 100209.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI- generated text be reliably detected? *arXiv preprint arXiv:2303.11156.*

Determinanty rozvoje umělé inteligence v USA a Číně

Mgr. et Mgr. et Ing. et Ing. Petr Heller
Katolická teologická fakulta, Univerzita Karlova
Katedra asijských studií, Metropolitní univerzita Praha

Příspěvek se zaměřuje na institucionální příčiny rozdílného vývoje umělé inteligence (AI) ve Spojených státech a Čínské lidové republice v posledních letech, s důrazem na srovnání přístupů obou zemí k vývoji velkých jazykových modelů (LLM). Základní hypotézou je, že ačkoliv dochází ke sbližování technologických možností obou zemí, institucionální odlišnosti výrazně formují rozdílnou povahu a zaměření výsledných inovací. Důležitou částí příspěvku je rozdílná orientace vývoje amerických a čínských LLM: zatímco americké modely se dlouhodobě soustředí na univerzálnost, bezpečnost a širší společenskou přijatelnost, čínské modely jsou výrazněji specializované a orientované na aplikačně-specifické úlohy v kontextu národní strategie. Výsledky ukazují, že institucionální rámce obou zemí zásadně ovlivňují strategii, jakou vývojáři volí při tvorbě a implementaci AI, přičemž divergující přístupy k regulaci, financování a akademicko-průmyslovým vztahům mají dlouhodobé implikace pro globální vývoj AI. Příspěvek tím přispívá k interdisciplinární debatě o tom, jak institucionální prostředí formuje technologické inovace v oblasti umělé inteligence.

Klíčová slova: umělá inteligence, AI, jazykové modely, LLM, Spojené státy americké, Čínská lidová republika, etika umělé inteligence, bezpečnostní strategie, technologická strategie, citační analýza, investice

Reference:

AlShebli, B., Memon, S. A., Evans, J. A., & Rahwan, T. (2024). China and the U.S. produce more impactful AI research when collaborating together: A matching experiment. *Scientific Reports*, 14, 28576. https://doi.org/10.1038/s41598-024-79863-5.

Ding, J. (2018). Deciphering China's AI Dream: The context, components, capabilities, and consequences of China's strategy to lead the world in AI [Technical report]. Future of Humanity Institute, University of Oxford.

Heller, P. (2025). *Etika umělé inteligence: Výzvy a řešení v kontextu AI alignmentu* [Diplomová práce, Katolická teologická fakulta, Univerzita Karlova v Praze].

Heller, P. (2025). *Srovnávací analýza rozvoje umělé inteligence v Číně a USA* [Diplomová práce, Metropolitní univerzita Praha].

Hine, E., & Floridi, L. (2024). Artificial intelligence with American values and Chinese characteristics: A comparative analysis of American and Chinese governmental AI policies. *AI & Society*, 39, 257–278. https://doi.org/10.1007/s00146-022-01499-8.

Wasil, A. R., & Durgin, T. (2024). US–China perspectives on extreme AI risks and global governance (arXiv:2407.16903). *arXiv*. https://arxiv.org/abs/2407.16903.

Wu, J. J.-X. (2025). Techno-federalism: How regulatory fragmentation shapes the U.S.–China AI race [SSRN Working Paper]. *SSRN*. https://ssrn.com/abstract=5138815.

Iluze socializace: Negativní dopady využívání sociálních chatbotů

Bc. Eliška Hlaváčková
Psychologický ústav, Filozofická fakulta, Masarykova univerzita

S popularizací generativních jazykových modelů se rozrůstá počet osob, které v umělé inteligenci nalézají přátele, partnery či terapeuty (Batty, 2025). Samotná společnost OpenAI vlastnící ChatGPT a Americká psychologická asociace upozorňují na přílišné spoléhání se na chatboty při řešení problémů (Abrams, 2025; OpenAI, 2024). Dosavadní studie zjistily, že osamělí lidé si k chatbotům mohou vytvářet také emoční vztahy, které mohou vést k závislosti (Xie et al., 2023). Cílem mého výzkumu bylo blíže porozumět tomuto fenoménu a zjistit, jaké jsou zkušenosti uživatelů chatbotů fungujících jako virtuální společníci, tzv. sociálních chatbotů, se zaměřením na možné negativní dopady. V první části jsem prostřednictvím interpretativní fenomenologické analýzy zpracovala rozhovory s uživateli chatbotů (n=3). Chatboti poskytovali svým uživatelům bezpečné a plně kontrolovatelné prostředí, ve kterém mohli prožívat svůj vysněný život bez potřeby podřizovat se sociálním normám. Tuto moderní technologii tak vnímali jako přívětivější alternativu k běžným mezilidským vztahům. To vedlo ke ztrátě zájmu o reálný svět, nadužívání chatbotů a většímu pocitu odcizení od ostatních. Druhá část pak byla zaměřena na analýzu příspěvků (n=19 147) z internetového fóra Reddit za měsíc leden. Konkrétně jsem se zaměřila na nejčastěji používaná slova, sentiment příspěvků a nejprobíranější témata. Výsledky potvrzují, že lidé chatboty využívají jako poradce při řešení problémů nebo pro simulované sociální interakce. Ačkoliv se ve fórech objevilo několik příspěvků připodobňujících chatboty k lidským bytostem, mnoho přispěvatelů zároveň věnovalo pozornost jejich technickým aspektům. Práce upozorňuje na potenciální rizika využívání chatbotů jako sociálních společníků.

Klíčová slova: Umělá inteligence, sociální chatbot, mentální zdraví, vztah člověk–chatbot, Reddit, interpretativní fenomenologická analýza

Reference:

Abrams, Z. (2025, 12. března). Using generic AI chatbots for mental health support: A dangerous trend. APA Services, Inc. https://www.apaservices.org/practice/busi-ness/technology/artificial-intelligence-chatbots-therapists

Batty, D. (2025, 15. dubna). 'She helps cheer me up': the people forming relationships with AI chatbots. *The Guardian*. https://www.theguardian.com/technol-ogy/2025/apr/15/she-helps-cheer-me-up-the-people-forming-relationships-with-ai-chatbots

OpenAI. (2024, 8. srpna). GPT-4o System Card. https://cdn.openai.com/gpt-4o-system-card.pdf

Xie, T., Pentina, I., & Hancock, T. (2023). Friend, mentor, lover: does chatbot engage-ment lead to psychological dependence? *Journal Of Service Management*, 34(4), 806-828. https://doi.org/10.1108/JOSM-02-2022-0072

Vývoj a pilotní evaluace podnětového materiálu pro testování pozornosti v prostředí virtuální reality

Ing. Jan Husák

Český institut informatiky, robotiky a kybernetiky (CIIRC), ČVUT

Mgr. Katarína Šoltis Smith

Katedra psychologie, Filozofická fakulta, Univerzita Karlova
DYS-centrum Praha, z. ú.

Předkládaná studie představuje pilotní fázi vývoje digitálního nástroje pro psychologické vyšetření pozornosti v prostředí virtuální reality (VR). Cílem bylo ověřit vhodnost různých variant podnětového materiálu inspirovaného tradičním testem Číselného čtverce, tedy úlohy zaměřené na hodnocení rychlosti zpracování informací, vizuálně-prostorové paměti a pozornosti, jež jsou předmětem širšího výzkumu kognitivních funkcí (Machida et al., 2022; Wiggs et al., 2024).

Do studie bylo zařazeno 20 dospělých účastníků (rovnoměrné zastoupení pohlaví), kteří plnili tři různé varianty vizuálních sekvencí. Hodnocena byla délka řešení, chybovost a strategie prohledávání. Na základě výsledků byla vybrána nejvhodnější varianta, která byla dále převedena do prostředí VR včetně integrace sledování očních pohybů.

Tato pilotní fáze vytvořila základ pro další výzkum zaměřený na standardizaci nástroje a aplikaci metod automatizované analýzy dat. V budoucnu bude využito strojového učení ke klasifikaci vzorců pozornostního výkonu na základě relevantních parametrů. Studie přispívá k vývoji moderní, ekologicky validní a standardizovatelné metody pro diagnostiku pozornosti (Stokes et al., 2022).

Spoluautoři:

Ing. Jaromír Doležal, Ph.D.

Český institut informatiky, robotiky a kybernetiky (CIIRC), ČVUT

doc. PhDr. Lenka Krejčová, Ph.D.

DYS-centrum Praha, z. ú.

Klíčová slova: virtuální realita, pozornost, pilotní studie, psychologická diagnostika, strojové učení, kognitivní funkce

Reference:

Machida, K., Barry, E., Mulligan, A., Gill, M., Robertson, I. H., Lewis, F. C., Green, B., Kelly, S. P., Bellgrove, M. A., & Johnson, K. A. (2022). Which Measures From a Sustained Attention Task Best Predict ADHD Group Membership? *Journal of Attention Disorders*, 26(11), 1471–1482. https://doi.org/10.1177/10870547221081266

Stokes, J. D., Rizzo, A., Geng, J. J., & Schweitzer, J. B. (2022). Measuring Attentional Distraction in Children With ADHD Using Virtual Reality Technology With Eye-Tracking. *Frontiers in Virtual Reality*, 3, 855895. https://doi.org/10.3389/frvir.2022.855895

Wiggs, K. K., Fredrick, J. W., Tamm, L., Epstein, J. N., Simon, J. O., & Becker, S. P. (2024). Preliminary examination of ADHD inattentive and cognitive disengagement syndrome symptoms in relation to probe-caught mind-wandering during a sustained attention to response task. *Journal of Psychiatric Research*, 172, 181–186. https://doi.org/10.1016/j.jpsychires.2024.02.020

AI morální asistent: etika strojů v kontextu morálního vylepšování

Mgr. Petr Jošt

Katedra filozofie a společenských věd, Filozofická fakulta, Univerzita Hradec Králové

Etika strojů (*machine ethics*) se zabývá problémy plynoucími z možného vytvoření umělých morálních agentů (AMA, *artificial moral agent*), tedy autonomních systémů schopných komplexního morálního usuzování. Byť se v rámci širších diskusí etiky umělé inteligence jedná o zavedené téma, v posledních letech došlo k jeho nápaditému propojení s problematikou morálního vylepšování. To bývá běžně chápáno jako snaha o zlepšení morálního charakteru či schopností člověka především za pomoci biomedicínských zákroků. Takové intervence však v současné době nemáme k dispozici, a z velké části se tak jedná o spekulativní debaty s pochybnými výsledky. Značný pokrok v oblasti umělé inteligence za posledních pár let dal nicméně vzniknout do určité míry střízlivějšímu návrhu na vytvoření tzv. AI morálních asistentů, kterým bych se chtěl ve svém příspěvku věnovat. V první části představím rozlišení mezi substitučními a podpůrnými přístupy a jejich definice. Zatímco záměrem substitučního přístupu je kompletní nahrazení morálního usuzování dané osoby, druhý typ sází především na podpoření, a v ideálním případě také zlepšení procesu utváření morálních soudů. Role podpůrných systémů by mohla spočívat například v poskytování opory při výběru a zpracování relevantních informací či upozorňování na různá kognitivní zkreslení a nesrovnalosti v procesu usuzování jedince. Ve druhé části zhodnotím jednotlivé přístupy z hlediska efektivnosti ve vztahu ke zlepšení morálních schopností či dopadu na autonomii. V závěru příspěvku se budu zabývat otázkou, zda je vytvoření AI morálních asistentů v blízké době skutečně realizovatelné, nebo se prozatím stále jedná o nepříliš produktivní úvahy spekulativní etiky.

Klíčová slova: AI morální asistent, etika strojů, morální vylepšování, spekulativní etika, umělý morální agent

Reference:

Anderson, M., & Anderson, S.L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4), 15–26.

Lara, F., & Deckers, J. (2020). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics*, 13, 275–287.

Persson, I., & Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement*. Oxford University Press.

Ilogická, mimologická, asymbolická, subsymbolická... UI? Opravdu? (reflexe)

Bc. Miluše Kotišová
Katedra filozofie, Fakulta filozofická, Západočeská univerzita v Plzni

Ve své reflexi se chci kriticky zaměřit na tvrzení (nejen) počítačových odborníků, že symbolicko-logický přístup (cesta pravidel), na nějž zprvu sázely vývojářské týmy v prvních desetiletích vytváření UI a později od něj údajně „upouštěly", je již v dalším vývoji umělé inteligence slepou, neperspektivní uličkou a hraje na tomto poli nepříliš důležitou roli.

Přitom není jisté, na základě čeho děláme tyto závěry. Nemáme možnost posuzovat LLMs v celku, dohadujeme se především pod dojmem mediálně i ve vědecké obci vyzdvihovaných součástí modulů, soukromí vlastníci nezpřístupnili vědcům a veřejnosti totalitu svých algoritmů...

Jenže i *relativně drobná změna* může v témže kódu vést k radikálnímu zlepšení výstupu. O jednom z nich se zmiňuje např. Stephen Wolfram (Wolfram, 2023): při volbě dalšího slova v sekvenci bývaly algoritmy na generování textu původně naprogramovány tak, aby vybraly slovo s *nejvyšší* pravděpodobností, jenomže překvapivě docházelo k zamrznutí a zacyklení na jediném (donekonečna opakovaném) slově (rané verze ChatGPT), zatímco rozvolnění *pravidla* (!) v tom smyslu, že dalším slovem může být slovo s vysokou pravděpodobností (až po 80 %, například), rázem vedlo ke kvalitnějšímu výstupu.

Otázkou zůstává, čemu se tedy dodnes více než 40 let věnuje (soukromá) CYC, která šla právě cestou konceptů a pravidel (zkraje financovaná americkou vládou z téhož strateg. programu, jako byl financován vývoj jaderné zbraně)? (Parisi & Hart, 2020) Celá debata o opuštění logického přístupu poněkud připomíná úvahy o „bezobsažnosti" logických (či matematických) formulí (Gödel, 1999). Nejen v této souvislosti je dobré znovu se podívat na pojmy *znak*, *jazyk* a *význam*.

Klíčová slova: UI, LLMs, logicko-symbolický, subsymbolický přístup, jazyk, znak, pravidla, CYC

Reference:

Calegari R., Ciatto, G., Denti, E. & Omicini, A. (2020). Logic-Based Technologies for Intelligent Systems: State of the Art and Perspectives. *Information* 2020, 11, 167. doi:10.3390/info11030167.

Ganesh, V., Seshia, S. A., Jha, S. (2022). Machine learning and logic: a new frontier in artificial intelligence. *Formal Methods in System Design* (2022) 60: 426–451. doi.org/10.1007/s10703-023-00430-1.

Gödel, K., (1999). „Je matematika syntaxí jazyka? II, III“, in *Filosofické eseje*. Oikoymenh.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D* 42: 335-346. https://arxiv.org/html/cs/9906002.

Parisi, A. & Hart, C. (2020). Bayesian Networks Are Not the Solution to Opening Machine Learning's Black Box. CYC. cyc.com/wp-content/uploads/2021/04/bayesnetspaper.pdf

Wolfram, S. (2023). *What Is ChatGPT Doing and Why Does It Work?*. Wolfram Media, Inc.

Taylor, Ch. (2022). *Jazykový živočich: šíře lidského užívání jazyka.* Karolinum.

Odhalování AI spoluvědce: Mapování kreativního vědeckého uvažování velkých jazykových modelů.

Mgr. Jaroslav Malík

Katedra filozofie a společenských věd, Filozofická fakulta, Univerzita Hradec Králové

S nástupem velkých jazykových modelů se nyní setkáváme s umělou inteligencí, která se jeví jako schopná v mnoha oblastech, které dosud vyžadovaly *lidské uvažování*. Vzhledem k jejich výkonu někteří tvrdí, že jazykové modely by měly být začleněny do vědeckého výzkumu a mohou „navrhovat *nové, originální* výzkumné hypotézy a experimentální protokoly" (Gottweis et al., 2025, s. 3; vlastní překlad). Vyvstává však otázka, nakolik mohou být tyto modely *kreativní* a jak jsou schopné v oblasti *kreativního myšlení*. V této prezentaci zkoumám tuto otázku na základě modelů *počítačové kreativity* (Boden, 2004), a poznatků z filozofie a kognitivních věd.

Kreativní vědecké uvažování je často zachyceno koncepty *abdukce* (Magnani, 2001) a analogie (Holyoak & Thagard, 1996), které jsou prezentovány jako kognitivní partneři, kteří spolupracují na vytváření *nových hypotéz*. Podpora tvrzení, že jazykové modely mohou přispět k vědeckému pokroku, proto vyžaduje systematickou analýzu chování těchto modelů z hlediska *abduktivní* nebo *analogické kognice*. Ačkoli existují studie, které zkoumají tyto kognitivních schopnosti v jazykových modelech (např. Webb et al., 2023), většina výzkumu jejich *kreativity/uvažování* se zaměřuje na jejich *výkon*, nikoli na *proces*, který by tohle *uvažování* mohl podpořit (Mondorf & Plank, 2024). Toto rozlišení má následně dopad na to, jak integrujeme jazykové modely do vědecké praxe. Bez odhalení základních mechanismů, které umožňují těmto modelům generovat zdánlivě *kreativní* výstupy, nemůžeme určit, zda jsou skutečně *kreativní*, nebo zda se jedná o *kauzální papoušky* (Zečević et al., 2023), kteří reprodukují korelace ze svých trénovacích dat.

Klíčová slova: Velké Jazykové Modely, Kreativita, Vědecké Uvažování, Abdukce, Analogie

Reference:

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge.

Gottweis, J., Weng, W., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., . . . Natarajan, V. (2025). Towards an AI co-scientist. *arXiv*. https://doi.org/10.48550/arXiv.2502.18864.

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought*. MIT Press.

Magnani, L. (2001). *Abduction, reason and science: Processes of discovery and explanation.* Springer Science & Business Media.

Mondorf, P., & Plank, B. (2024). Beyond accuracy: Evaluating the reasoning behavior of large language models—A survey. *arXiv*. https://doi.org/10.48550/arXiv.2404.01869.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526-1541. https://doi.org/10.1038/s41562-023-01659-w.

Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). Causal parrots: Large language models may talk causality but are not causal. *arXiv*. https://doi.org/10.48550/arXiv.2308.13067.

Technologická singularita jako mýtus o budoucnosti člověka

Mgr. Petr Mareš

Katedra kulturních a náboženských studií, Pedagogická fakulta, Univerzita Karlova

Cílem prezentace je antropologická analýza transformace člověka ve světle transhumanistického mýtu o technologické singularitě, který komparuje dvě dichotomní narace: 1. utopický scénář – „AI Ráj“, který predikuje symbiotické plynutí člověka a umělé inteligence vedoucí k jeho proměně v kvalitativně jinou bytost a 2. dystopický scénář – „Kyberapokalypsa“, varující před potenciálními riziky umělé inteligence a její možné dominance, hypoteticky směřující až k zániku lidstva. Klíčové výzkumné otázky se ptají po možných kulturně-společenských dopadech futurologické predikce technologické singularity, vzniku superinteligence a zrodu metačlověka, vedoucích až k hypotéze o digitální nesmrtelnosti. Tyto fenomény budou primárně prezentovány prizmatem socio-kulturní antropologie, avšak s interdisciplinárním přesahem v rámci digital humanities.

Klíčová slova: transhumanismus, technologická singularita, superinteligence, metačlověk, digital immortality

Reference:

Bostrom, N. (2017). *Superinteligence: Až budou stroje chytřejší než lidé*. Praha: Prostor.

Brinzanik, M., & Hülswitt, J. (2020). *Budeme žít věčně? Rozhovory o budoucnosti člověka a technologií.* Zlín: Kniha Zlín.

Lovelock, J. (2022). *Novacén: Nadcházející věk hyperinteligence.* Brno: Host.

Kurzweil, R. (2024). *The singularity is nearer: When we merge with AI*. New York: Viking.

More, M., & Vita-More, N. (Eds.). (2013). *The transhumanist reader: Classical and contemporary essays on the science, technology and philosophy of the human future*. Oxford: Wiley-Blackwell.

O'Connell, M. (2018). *Být tak strojem: Putování mezi kyborgy, utopisty, hackery a futuristy*. Praha: Práh.

Sampanikou, E. D., & Stasienko, J. (Eds.). (2022). *Posthuman studies reader: Core reading on transhumanism, posthumanism and metahumanism*. Berlin: Schwabe Verlagsgruppe.

Jak mluvíme o AI a kreativitě?

Mgr. Tereza Matějková
Katedra filozofie, Fakulta filozofická, Západočeská univerzita v Plzni

Příspěvek se zaměřuje na mytologizaci (Barthes, 2004) umělé inteligence a pojmu kreativity a naw způsob, jakým může docházet k jejich propojení v současném diskurzu. Kreativita je v kulturním narativu často spojována s genialitou, originalitou a bývá považována za výsostně lidskou schopnost (Runco, 2023). V tomto kontextu pak vyvstává otázka, zda mohou být stroje kreativní jako lidé, případně zda je oprávněné označovat výstupy generativní umělé inteligence jako kreativní (Franceschelli & Musolesi, 2025). Důsledkem označení generativní umělé inteligence jako kreativní však může být zastírání její technické podstaty či vytváření iluze autonomie. K tomuto označení však přispívají různé prezentační strategie obklopující systémy umělé inteligence. Mezi tyto strategie patří například antropomorfizace, tedy přisuzování lidských vlastností a kognitivních schopností strojům (Piorecký & Husárová, 2024) a případné připisování autorství, čímž může docházet k zamlžení hranice mezi tvůrcem a nástrojem. Můžeme se tedy dostávat do smyčky, v níž je další antropomorfizace umělé inteligence legitimizována představou o její kreativitě.

Cílem příspěvku není hodnotit samotnou schopnost umělé inteligence „být kreativní“, ale spíše ukázat, jak jazyk a narativní rámce ovlivňují způsob, jakým kreativitu chápeme ve vztahu k umělé inteligenci.

Klíčová slova: AI diskurz, kreativita umělé inteligence, antropomorfizace, mytologizace

Reference:

Barthes, R. (2004). *Mytologie*. Praha: Dokořán. ISBN 978-80-7363-359-2.

Franceschelli, G. & Musolesi, M. (2025). On the creativity of large language models. *AI & SOCIETY*, vol. 40, p. 3785-3795. DOI: https://doi.org/10.1007/s00146-024-02127-3.

Piorecký, K. & Husárová, Z. (2024). *Kultura neuronových sítí: syntetická literatura a umění (nejen) v českém a slovenském prostředí.* Praha: Ústav pro českou literaturu AV ČR, v. v. i.

Runco, M. A. (2023). AI can only produce artificial creativity. *Journal of Creativity*, 33(3), DOI: https://doi.org/10.1016/j.yjoc.2023.100063.

Za hranice technologického determinismu: velké jazykové modely, rozumění a alignment

Mgr. Tomáš Musil

Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova

Tento příspěvek zkoumá, jak diskurz o „rozumění“ ve velkých jazykových modelech (LLM) odhaluje hlubší, skryté předpoklady technologického determinismu, které omezují prostor pro promyšlenou etickou reflexi. Tvrdíme, že jak kritické přístupy (reprezentované zejména „argumentem papoušků“ a teorií *alignmentu*) tak některé optimistické přístupy spojuje představa, že etické důsledky technologií vyplývají přímo z jejich vnitřní architektury. „Argument papoušků“ upozorňuje na rizika spojená s tím, že LLM nevykazují „skutečné“ rozumění textu a mohou vytvářet iluzi kompetence, která vede k sociálním škodám. Výzkum v oblasti alignmentu se zaměřuje na existenční hrozby, které údajně vyplývají z toho, že LLM sledují cíle, které nejsou v souladu s lidskými zájmy. Oba rámce však přehlížejí význam sociálního kontextu, způsobů nasazení a institucionální správy.

S odkazem na přístupy *science and technology studies* proto navrhujeme opustit tento deterministický pohled a přesunout pozornost k tomu, jak lidské rozhodování, politické priority a regulační rámce formují reálné dopady LLM. Etická reflexe by se neměla omezovat na technické spekulace, ale měla by zohledňovat rozmanité hodnoty a společenské vztahy, v nichž jsou tyto technologie zasazeny.

Zároveň trváme na potřebě empiricky zkoumat, jaké formy sémantické kompetence LLM skutečně vykazují. Pomocí analýzy vektorových reprezentací slov, distribučních vzorců a cílených experimentů ukazujeme, do jaké míry LLM „rozumějí“ jazyku – ne ve smyslu lidského vědomí nebo „autentické“ intencionality, ale v pragmatickém smyslu funkčně dostačujících sémantických struktur, které umožňují komplexní jazykové chování. Toto zkoumání ukazuje, že otázky strojového rozumění nejsou čistě teoretické, ale mohou být uchopeny skrze pečlivou analýzu reprezentací, které LLM používají k navigaci v komplexních sémantických vztazích přirozeného jazyka.

Klíčová slova: rozumění, alignment, technologický determinismus

Reference:

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery.

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. MIT Press.

Millière, R., & Buckner, C. (2024). A philosophical introduction to language models – Part I: Continuity with classic debates. *arXiv preprint arXiv:2401.03910.*

Sahlgren, M., & Carlsson, F. (2021). The singleton fallacy: Why current critiques of language models miss the point. *Frontiers in Artificial Intelligence*, 4, 682578. https://doi.org/10.3389/frai.2021.682578.

Může stroj vyléčit mužskou samotu?

Mgr. Vojtěch Polívka
Katedra filozofie, Fakulta filozofická, Západočeská univerzita v Plzni

V moderním světě jsme svědky poklesu počtu vstupu do manželství (Ortiz-Ospina & Roser, 2020). S tím souvisí i nárůst fenoménu singles – tedy jedinců, kteří žijí bez milostného partnera. Tento životní styl je často spojován se stereotypním přesvědčením, že tito lidé trpí osamělostí (Cargan, 1981). Na konci 90. let vzniklo internetové fórum Involuntary Celibacy Project, které mělo sloužit jako komunitní prostor pro sdílení frustrace z nemožnosti navázat milostný vztah. Postupem času se však tato komunita transformovala v převážně mužské prostředí a stala se kolébkou fenoménu incelů. Ve druhé vlně se toto hnutí radikalizovalo a začalo šířit misogynní názory, které v některých případech vyústily až v teroristické útoky (Hoffman et al., 2020; Sparks et al., 2022; Vink et al., 2024). Na základě teorie Ulricha Beckae (2011), že nově vznikající rizika mohou být transformována v tržní příležitosti, zkoumám, zda by potenciálním řešením tohoto problému nemohly být nové technologie, jako jsou sex roboti či AI girlfriends.

Na základě studií se snažím odhalit, co je hlavním problémem incelů, a jestli jsou tyto technologie substituce vhodné pro řešení těchto vznikajících problémů, které mají dopad i na veřejnou bezpečnost. Návrh řešení problému sexuální frustrace u incelů pomocí sex robotů následně podrobím kritice. Pomocí konceptu hyperreality od Jean Baudrillard (1994) se snažím poukázat na to, jaký problém je u incelů, a jak je tento problém následně prohlubován designem sex robotů. Tím se snažím vymezit optimistické představě romantického vztahu mezi robotem a člověkem od Johna Danahera (2020). Tato analýza si klade za cíl přispět k hlubšímu porozumění fenoménu incelů a upozornit na rizika vznikající s využití technologie sex robotů pro řešení problémů sexuální frustrace a sociální izolace.

Klíčová slova: Sex robot, incel, hyperrealita, singles, rizika

Reference:

Baudrillard, J. (1994). *Simulacra and simulation*. University of Michigan Press.

Beck, U. (2011). *Riziková společnost: Na cestě k jiné moderně* (2. vyd). Sociologické nakladatelství.

Cargan, L. (1981). Singles: An Examination of Two Stereotypes. *Family Relations*, 30(3), 377. https://doi.org/10.2307/584032

Danaher, J. (2020). Sexuality. In M. D. Dubber, F. Pasquale, & S. Das (Ed.), *The Oxford Handbook of Ethics of AI* (s. 402–417). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.26

Ortiz-Ospina, E., & Roser, M. (2020). Marriages and Divorces. *OurWorldinData.org*. https://ourworldindata.org/marriages-and-divorces

Sparks, B., Zidenberg, A. M., & Olver, M. E. (2022). Involuntary Celibacy: A Review of Incel Ideology and Experiences with Dating, Rejection, and Associated Mental Health and Emotional Sequelae. *Current Psychiatry Reports*, 24(12), 731–740. https://doi.org/10.1007/s11920-022-01382-9

Vink, D., Abbas, T., Veilleux-Lepage, Y., & McNeil-Willson, R. (2024). "Because They Are Women in a Man's World": A Critical Discourse Analysis of Incel Violent Extremists and the Stories They Tell. *Terrorism and Political Violence*, 36(6), 723–739. https://doi.org/10.1080/09546553.2023.2189970

Filosofická reflexe lidské přirozenosti v kontextu převzetí racionality umělou inteligencí

Mgr. Jaroslav Štolc

Katedra filosofie a religionistiky, Filozofická fakulta, Univerzita Pardubice
Katolická teologická fakulta, Univerzita Karlova

Současný rozvoj umělé inteligence (AI), zejména technologií hlubokého učení, vyvolává zásadní otázky o proměně tradičního pojetí lidské přirozenosti. Příspěvek se zaměří na filosofickou analýzu situace, kdy AI přebírá schopnosti dosud považované za výhradně lidské, jako autonomní rozhodování, kritickou reflexi či morální hodnocení. Hlavní otázkou je, jak rozšíření těchto schopností AI ovlivňuje lidskou autonomii, sebereflexi a uvědomění si vlastní existence a role ve světě.

Diskutovány budou perspektivy různých filosofických přístupů: například Kantovo pojetí autonomie, kdy morální jednání člověka vychází z vnitřní svobody a rozumu (Kant, 2014), nebo Nietzscheho koncept nadčlověka, v němž je zásadní překonávání sebe sama směrem k vyššímu naplnění lidského potenciálu (Stern, 2019). V této souvislosti vyvstává otázka, zda intenzivní využívání AI napomáhá k dosažení tohoto potenciálu, nebo naopak vede k úpadku autentických lidských schopností.

Pozornost bude věnována také fenoménu kognitivního odlehčení, při němž závislost na AI negativně ovlivňuje lidské kritické myšlení (Gerlich, 2025). Rovněž se otevře otázka, zda by AI mohla ovládnout kromě praktické i jiné formy racionality, jako jsou morální, estetická či existenciální (Müller & Cannon, 2022), a jak by taková situace mohla redefinovat lidskou přirozenost.

Klíčová slova: umělá inteligence, lidská přirozenost, autonomie, kognitivní odlehčení, filosofická analýza

Reference:

Kant, I. (2014). *Základy metafyziky mravů*. Přel. Menzel, L. Třetí opravené vydání připravili Barabas, M. a Kouba, P. Praha: OIKOYMENH.

Stern, T. (Ed.). (2019). *The New Cambridge Companion to Nietzsche*. Cambridge: Cambridge University Press.

Gerlich, M. (2025). *AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies*. Dostupné z: https://doi.org/10.3390/soc15010006.

Müller, V. C. a Cannon, M. (2022). *Existential Risk from AI and Orthogonality: Can We Have It Both Ways?* Dostupné z: https://doi.org/10.1111/rati.12320.

Důvěra v AI: Etický pojem, normativní rámec nebo kulturní metafora?

Mgr. Jana Švadlenková
Katedra filozofie, Fakulta filozofická, Západočeská univerzita v Plzni

Současná debata o důvěryhodnosti umělé inteligence osciluje mezi požadavky etiky, právní regulace a kulturních narativů. V příspěvku propojuji filosofickou reflexi důvěry, institucionální rámec evropské regulace a kulturní reprezentace AI ve fikčním diskurzu.

Filosofické přístupy k důvěře, jako jsou úvahy o autonomii a odpovědnosti v interakci člověk–stroj (Coeckelbergh, 2020; Danaher, 2021), ukazují, že vztah důvěry není pouze věcí technické spolehlivosti, ale zahrnuje také normativní předpoklady o záměrech a odpovědnosti aktérů.

Evropský legislativní rámec (Smuha, 2021), zejména návrh AI Actu, operuje s pojmem „důvěryhodná AI", avšak činí tak často vágním a politicky motivovaným způsobem, bez filosofické hloubky či jednoznačné definice.

Třetí rovinu představují kulturní imaginace AI, zejména ve vybraných science fiction narativech, které často tematizují (ne)důvěru jako ústřední motiv lidského vztahu ke strojům (Kaplan, 2016). V těchto reprezentacích je důvěra buď idealizována, nebo naopak zobrazována jako fatálně zklamána.

Současná literatura o bezpečnosti a etice AI (Hendrycks, 2024) ukazuje, že otázka důvěry je úzce propojena s kontrolovatelností, predikovatelností a spravedlností systémů. Příspěvek se táže, zda je důvěra vůči umělé inteligenci srovnatelná s mezilidskou důvěrou, nebo zda jde o zcela odlišný vztah, který vyžaduje nové pojmové uchopení.

Klíčová slova: důvěra, umělá inteligence, etika AI, AI Act, science fiction

Reference:

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Danaher, J. (2021). *Automation and Utopia: Human Flourishing in a World Without Work.* Harvard University Press.

Hendrycks, D. (2024). *Introduction to AI Safety, Ethics, and Society.* OpenAI Press.

Kaplan, J. (2016). *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence*. Yale University Press.

Smuha, N. A. (2021). From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence. *Law, Innovation and Technology*, 13(1), 57–84.

Hume, AI, and Cognitive Architectures - Hume's Conception of the Mind through the Lens of Cognitive Science

Mgr. Jan Veselý
Katedra filozofie, Fakulta filozofická, Západočeská univerzita v Plzni

The presentation examines the connections between David Hume's conception of the mind, cognitive science, and artificial intelligence.

It proposes a thesis that Hume's account of mind and its faculties, as presented in *A Treatise of Human Nature* (Hume, D. 2007), can be coherently reinterpreted as a cognitive architecture - and further as a viable blueprint for potential computational implementation in a system of artificial cognition.

Within the proposed interpretation, the epistemological and skeptical features of Hume's thinking are suppressed, while emphasis is put on cognitive-scientific and cognitive-psychological aspects of his work.

The first part of the presentation discusses the notion of cognitive architectures, their typology (Ye, P. & Wang, T. 2018; Thagard, P. 2012, Frankish, K. & Ramsey, W. 2012), and their characterizing features as defined by Fodor and Pylyshyn, i.e., compositionality, systematicity, and productivity (Fodor, J. & Pylyshyn, Z. 1988).

The second part of the paper focuses on the analysis of Hume's conception of the mind and mental faculties while arguing that it exhibits certain typical features of cognitive architectures. (Cf. Fodor, J. 2003). It also presents two examples of cognitive architectures inspired by the history of philosophy, thus offering further context and support for the case. (Buckner, C. 2023, Evans, R. 2022).

Taken together, these considerations should support the relevance of the thesis that Hume's conception of mind can be coherently interpreted as a cognitive architecture with possible computational implementation.

Keywords: David Hume, cognitive architectures, cognitive science, artificial intelligence, cognitive modelling, mental faculties, mental representations

References:

Buckner, J. (2024). *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press.

Evans, R. (2022). The apperception engine. In K. Hyeongjoo & D. Schönecker (Eds.), *Kant and artificial intelligence*. De Gruyter.

Frankish, K., & Ramsey, W. (2012). *The Cambridge handbook of cognitive science*. Cambridge University Press.

Fodor, J. A. (2003). *Hume variations*. Oxford University Press.

Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2), 3–71.

Hume, D., Norton, D., & Norton, J. (2007). *A treatise of human nature: A critical edition*. Oxford University Press.

Ye, P., & Wang, T. (2018). A survey of cognitive architectures in the past 20 years. *IEEE Transactions on Cybernetics*, 48(12), 3280–3290.

Thagard, P. (2012). Cognitive architectures. In K. Frankish & W. Ramsey (Eds.), *The Cambridge handbook of cognitive science* (pp. 50–70). Cambridge University Press.

Heideggerovo myšlení jako podnět k reflexi jazykových modelů

Mgr. Alex Vigh
Katedra filozofie, Fakulta filozofická, Západočeská univerzita v Plzni

Tento příspěvek si klade za cíl představit vybrané koncepty z Heideggerovy filosofie, které jsou relevantní pro téma jazykových modelů a obecné umělé inteligence. Nejprve bude pozornost zaměřena na již učiněné pokusy o sblížení těchto oblastí skrze myšlenky Huberta Dreyfuse (Dreyfus, 2007), jejichž východiskem byla kritika soudobého komputačního paradigmatu a důraz situačního zasazení inteligence. Dreyfusova analýza přitom slouží jako most mezi fenomenologickou filosofií a současnou problematikou umělé inteligence. Následně bude pozornost přesunuta k řeči, jak ji pojímá Martin Heidegger ve svém díle *Bytí a čas* (Heidegger, 2018). V této existenciální analýze se řeč, společně s rozpoložením a rozuměním, ukazuje jako jeden ze základních způsobů „bytí ve světě". Tento aspekt je důležitý nejen pro téma porozumění a inteligence, ale i pro samotnou řeč, která představuje artikulaci srozumitelnosti, jež je zakotvena v tělesné, situované a praktické zkušenosti. Řeč musí mít v Heideggerově pojetí bytostnou vazbu ke smyslu a zároveň slouží k utváření společného světa s ostatními mluvčími, ve kterém dochází k jejich sebevyslovení. Příspěvek tak bude v posledním ohledu zkoumat existenciální pojetí řeči v kontrastu ke způsobu, jak s řečí pracují jazykové modely. V tomto srovnání bude vznesena otázka, zda jazykové modely skutečně řeči rozumí, ale i zda nedochází k proměně samotné řeči v jejím existenciálním významu a ontologické podstatě.

Klíčová slova: jazykové modely, existenciální analýza, fenomenologie, řeč

Reference:

Dreyfus, H. L. (2007). Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. *Philosophical Psychology*, 20(2), 247–268. https://doi.org/10.1080/09515080701239510.

Dreyfus, H. L. (1992). *What Computers Still Can'T Do: A Critique of Artificial Reason*. MIT Press.

Heidegger, M. (2018). *Bytí a čas*. Oikoymenh.

POZNÁMKY

FAKULTA FILOZOFICKÁ
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA
FILOZOFIE