Jakub Dvořák

AI Safety Researcher

MFF UK · Prague

Scheming & deception detection in language models

CONTACT

hi@kubadvorak.com
+420 722 020 622
Prague, Czech Rep.
jakubdvorak-ai
github.com/Majny
@jakubdvorak_ai
kubadvorak.substack.com

LANGUAGES

Czech ●●●●●
English ●●●●○

CORE SKILLS

Research methods
Behavioral evaluation, in-context scheming benchmarks, alignment-faking classifiers, chain-of-thought analysis, paper replication

Evaluation toolkit
inspect-ai, Apollo's evaluation harness, Anthropic alignment-faking classifier, custom prompt suites, statistical testing

ML / DL
PyTorch, transformers, attention mechanisms, fine-tuning, RLHF, reasoning models (DeepSeek-R1, Qwen QwQ)

Programming
Python, TypeScript, Kotlin, C++ (modern), C, Bash

Engineering
Microservices, PostgreSQL, Docker, JWT/RS256 auth, Linux (Arch), CI/CD

AWARDS

Red Bull Basement '26
Czech Republic National Top 10

BUDETO Studio
Backed Founder · €50k angel ticket

SCIO Mathematics
98th percentile

Podnikni to!
CTU Entrepreneurship Program

PROFILE

AI safety researcher with a strong systems engineering foundation. Currently focused on **scheming and deception detection** in frontier language models — particularly in-context scheming behaviors of open-weight reasoning models (DeepSeek-R1, Qwen QwQ) and how those behaviors generalize across languages. Finishing my Bachelor's in Computer Science at the Faculty of Mathematics and Physics, Charles University (MFF UK) and starting a Master's in Artificial Intelligence in September 2026, with research direction at ÚFAL.

Bachelor's thesis is a multisig Bitcoin wallet on Android with M-of-N support (BIP-48/67/174), Trezor integration, PSBT distribution between cosigners, and a 140-test automated suite covering BIP-encoding invariants, race-safe token rotation, and audit-trail correctness. Code, replication notes, and findings published openly. Concurrently building an AI infrastructure venture; research is the primary focus through 2026–2028.

RESEARCH INTERESTS

- **In-context scheming and deception detection** in frontier reasoning models — behavioral evaluation of scheming capabilities, alignment faking, and sandbagging across model families (Apollo Research benchmarks, Anthropic alignment faking methodology).
- **Cross-lingual scheming evaluation** — whether scheming behaviors and chain-of-thought deception generalize across languages, and what that implies for safety of multilingual deployment. An underexplored area; US-based research is constrained by language access.
- **Chain-of-thought monitorability** in open-weight reasoning models — when CoT is faithful versus post-hoc rationalization, and how monitoring techniques degrade across languages and model families.
- **Practical alignment for production deployment** — the context I encounter daily building AI infrastructure, where current alignment methods are insufficient and where rigorous evaluation would help.

IN PROGRESS

Cross-lingual in-context scheming detection in open-weight reasoning models Apollo + Anthropic methodology · DeepSeek-R1, Qwen QwQ · Czech extension

Replicating Apollo Research's "Frontier Models Are Capable of In-Context Scheming" (Meinke et al., December 2024) and Anthropic's "Alignment Faking Revisited" (April 2025, open source) on open-weight reasoning models. Extending with a cross-lingual analysis: translating the scheming evaluation suite into Czech, running on R1 and QwQ in English, Czech, and bilingual conditions, and measuring whether scheming rates and chain-of-thought faithfulness generalize across languages. Code, notes, and findings will be published openly on GitHub and as a LessWrong write-up.

EDUCATION

Charles University, Faculty of Mathematics and Physics (MFF UK) Sep 2026 — 2028

Master's in Artificial Intelligence · planned

Continuing at MFF UK after the Bachelor's. Research direction: scheming and deception detection in frontier reasoning models, cross-lingual chain-of-thought analysis. Prospective advisor: ÚFAL faculty.

Charles University, Faculty of Mathematics and Physics (MFF UK) 2023 — 2026

Bachelor of Science, Computer Science

Bachelor's thesis: *Bitcoin Wallet for Advanced Users*, supervised by RNDr. Filip Zavoral, Ph.D. Defence 18 Jun 2026; final state exam 21 Jun 2026.

Jakub Dvořák

AI SAFETY · MFF UK

Curriculum vitae — research variant.
Audience: MATS, Anthropic Fellowship, Apollo Research, Goodfire, ARENA, AI Safety Camp, PhD applications.

hi@kubadvorak.com

kubadvorak.com

University of Chemistry and Technology, Prague (VŠCHT) 2022 — 2023
Bachelor's, Synthesis and Production of Pharmaceuticals (transferred)

Gymnázium Mladá Boleslav 2018 — 2022
Czech maturita*, eight-year selective gymnasium program*

SELECTED PROJECTS

Bitcoin Wallet for Advanced Users Bachelor's thesis · MFF UK · Jun 2026

Android application for advanced Bitcoin asset management: M-of-N multisig (BIP-48, BIP-67), coin control, Trezor hardware-wallet integration via Trezor Connect, PSBT distribution between cosigners (BIP-174). Backend: Kotlin/Ktor microservices on PostgreSQL with Docker Compose. Frontend: Jetpack Compose. Verified on Bitcoin testnet with a complete 2-of-3 multisig transaction. Includes a 140-test automated suite (5 modules) covering PSBT wire format, BIP-32/67 derivation, BIP-380/383 descriptor parsing, race-safe single-use refresh-token rotation (verified concurrently across 8 threads), and audit-trail integrity. Several integration issues with Trezor firmware (BIP-67 child-level ordering, `PSBT_IN_NON_WITNESS_UTXO` for SegWit inputs) were diagnosed and fixed; documented in chapter 4.4 of the thesis.

Boletiqo AI company platform · past project (2026)

Co-founded with Lukáš Hellesch. Autonomous AI agent teams (CEO, CTO, CMO, CFO + specialists) that plan, ship, sell, and close. Red Bull Basement '26 Czech Republic national top 10. Building this surfaced concretely how unreliable LLM agents become at the edge of their capability — agents misrepresenting intermediate results, pursuing instrumental subgoals that diverged from the spec, and producing convincing post-hoc justifications. The experience informed my pivot to scheming and deception detection in frontier reasoning models, which is now my primary research focus.

RISC-V operating-system kernel team-losOS · 3-person team · NSWI200

Built bottom-up: console output, bump-pointer allocator, thread scheduling, interrupt handling, virtual memory with page-table protection, system calls, custom minimal libc, multi-process userspace with stack-overflow detection. Tested on the MSIM simulator with a CI suite covering every milestone.

Ant Colony Simulator Unity · C# · ShaderLab/HLSL

Multi-colony foraging simulator demonstrating emergent intelligence from local rules. Pheromone trails with diffusion and decay; per-colony genome; parallel multi-colony execution; sandbox editor for map design. PlantUML architecture documentation.

EXPERIENCE

Co-founder & CTO of an early-stage AI venture Mar 2026 — present
Stealth · Prague

- Building an AI agent runtime and inference infrastructure for a hosted product. Personally responsible for: agent-to-agent negotiation protocol, confidence-scored memory across categories, LLM-driven conversation engine, third-party API integrations.
- Backed by BUDETO Studio (€50k angel ticket).
- This is a useful counterweight to academic research — it forces me to take AI deployment realities seriously and puts me in the room when models actually break.

Kreedl · Prague, Czech Republic Jan 2026 — Jun 2026
Software Engineer

- Owned the end-to-end design and implementation of *Kreedl Intelligence*, an AI-powered pipeline for automated pitch-deck analysis: data ingestion, model orchestration, output ranking.

Jakub Dvořák

AI SAFETY · MFF UK

Curriculum vitae — research variant.
Audience: MATS, Anthropic Fellowship, Apollo Research, Goodfire, ARENA, AI Safety Camp, PhD applications.

PUBLICATIONS & PUBLIC WRITING

- *Forthcoming, 2026.* “Cross-Lingual In-Context Scheming Detection in Open-Weight Reasoning Models.” — LessWrong / personal blog write-up.
- *Forthcoming, 2026.* “From Systems Engineering to AI Safety Research.” — Substack post.

kubadvorak.substack.com

VENTURES

Appnestiq — Co-founder 2025 — 2026

First entrepreneurial venture, with Lukáš Hellesch. AI consulting and custom AI implementation services. Shut down without paying customers; the legal entity now serves as the operating shell for the current venture’s infrastructure.

hi@kubadvorak.com
kubadvorak.com