



Jakub Dvořák

Systems Engineer → AI Safety

MSc AI (from 2026) · MFF UK · Prague

🔗 Cross-lingual oversight
of reasoning 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 focus

Chain-of-thought monitorability, control-style deception evals, cross-lingual robustness of oversight, experiment design & eval statistics

Eval stack (building in public)

inspect-ai harnesses, vLLM serving of open-weight reasoning models (gpt-oss, Qwen3), LLM-as-judge pipelines, bootstrap & paired testing

ML foundations

Deep-learning coursework at MFF UK (transformers, attention); PyTorch in progress; LLM APIs in production (Claude, Gemini, GPT)

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 pre-seed

SCIO Mathematics

98th percentile

Podnikni to!

CTU Entrepreneurship Program

PROFILE

Systems engineer and founder moving into AI-safety research. My question: **when chain-of-thought oversight of a language model fails outside English, is it the monitor or the model that fails?** The answer decides between a cheap translation-layer fix and a fundamental limit of CoT monitoring — and it is measurable with a careful control-style experiment, which I am building openly. Finishing my Bachelor's in Computer Science at the Faculty of Mathematics and Physics, Charles University (MFF UK); starting a Master's in Artificial Intelligence there in October 2026.

What I bring today is engineering execution: production systems shipped as founder/CTO, plus low-level systems work — a RISC-V kernel, and a multisig Bitcoin wallet with a 140-test automated suite (Bachelor's thesis). The research toolkit is the part I am deliberately building in public, with the first project itself. Research is the primary focus through 2026–2028.

RESEARCH INTERESTS

Theme: **multilingual robustness of AI oversight** — safety tooling is built and tested overwhelmingly in English; I study what happens to it everywhere else.

- **Cross-lingual chain-of-thought monitorability** — decomposing why monitors miss deceptive reasoning outside English: monitor-side failure (recoverable by translate-then-monitor) versus model-side signal loss. Directly extends the 2026 cross-lingual fragility results, whose authors explicitly defer the agentic setting to future work.
- **Language mixing as a covert channel** — reasoning models code-switch naturally mid-thought; whether that hides scheme-relevant reasoning from trusted monitors, and how cheaply a language-normalization defense closes the gap.
- **Evaluation awareness across languages** — whether latent and verbalized “this is a test” signals are language-bound, which reframes how much non-English safety evals can be trusted.
- **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

Lost in the Monitor: is cross-lingual CoT oversight failure monitor-side or model-side?

control-style eval · Inspect · gpt-oss-20b, Qwen3-32B · autumn 2026

A monitor-language × reasoning-language evaluation matrix over agentic deception transcripts with *programmatically* ground truth (instructed-covert code-sabotage scored by hidden tests), plus the deciding arm: translating the chain of thought to English before monitoring and measuring how much detection recovers. Human-verified Czech translations are the data layer most labs cannot produce in-house; monitors include an LLM judge, a translate-then-monitor pipeline, a weak trusted model, and a keyword baseline, compared by AUC with paired testing and pre-registered thresholds. Timeline: harness and translation packs summer 2026; two-week MVP in September; public research note early October; full write-up with open code and all transcripts by end of November 2026.

EDUCATION

Charles University — MFF UK

Oct 2026 — 2028

Master's in Artificial Intelligence · planned

Continuing at MFF UK after the Bachelor's. Research theme: multilingual robustness of AI oversight — CoT monitorability, deception evals, cross-lingual evaluation. Prospective supervision at ÚFAL.

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 move into research on oversight and deception in LLM agents, which is now my primary focus.

RISC-V operating-system kernel team-losOS · 3-person team · NSW1200

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 Zoplio Mar 2026 — present

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 pre-seed).
- › 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.