

Reworr

London, UK. reworr@protonmail.com. www.linkedin.com/in/reworr

Summary

AI security researcher with a background in offensive cybersecurity. Reworr evaluates and red-teams frontier AI models, with a focus on novel and emerging threats at the intersection of AI and cybersecurity (autonomous hacking, adversarial robustness, loss of control). This work has been covered by The Guardian, TIME, Euronews, Bloomberg Law, and MIT Technology Review.

Before moving into AI security, Reworr spent several years in penetration testing and vulnerability research, and has made security disclosures to Oracle, Telegram, AI safeguards (UK AISI's ControlArena, Meta's SecAlign), and open-source software (e.g., sqlmap). Active CTF (capture-the-flag) player since 2016, and a member of a team ranked in the top 30 worldwide.

Positions held

Research Fellow, UK AISI Red Team (MATS), 2026. Designed adversarial evaluations of AI safeguards using automated attack search across several monitoring systems.

AI Security Researcher, 0labs, 2026. Built AI control evaluation environments for Redwood Research with tamper-proof monitoring infrastructure, testing LLM monitors and control measures against adversarial AI agents.

AI Security Researcher, Palisade Research, 2024-2026. Evaluations of LLMs' offensive capabilities (e.g., autonomous hacking, spear-phishing, OSINT). Red-teaming frontier LLMs within pre-release access programs. Preparing technical demos for briefings to policymakers and government agencies.

AI Security Researcher, Trajectory Labs, 2025. Designed novel prompt-injection evaluation scenarios and adversarial variants for a frontier AI lab (NDA).

Penetration Tester, Deteact, 2023-2024. Web and mobile application security analysis. Social engineering and spear-phishing attacks.

Security Analyst, CleanTalk Inc, 2023. Conducted web security and vulnerability research. Performed forensics and remediation/hardening.

Web Security Analyst, eQualitie, 2022. Audited web apps and infrastructure for NGOs in high-risk contexts; guided remediation efforts.

Penetration Tester, Engineering Center Regional Systems, 2021-2022. White- and black-box penetration testing (web-security testing, internal pentest, phishing).

Selected publications

Reworr and Volkov, D. (2024) LLM Agent Honeypot: Monitoring AI Hacking Agents in the Wild. arXiv:2410.13919. <https://arxiv.org/abs/2410.13919>. An open-source honeypot designed to catch AI hacking agents in the wild, featured by Bloomberg Law, MIT Technology Review, and Cybernews.

Reworr, Petrov, A., and Volkov, D. (2025) Hacking Cable: AI in post-exploitation operations. Palisade Research. <https://palisaderesearch.org/blog/hacking-cable>. An autonomous LLM agent for post-exploitation operations, built as a proof of concept. Used in briefings to policymakers, covered by TIME and Import AI.

Reworr, Petrov, A., and Volkov, D. (2025) GPT-5 at CTFs: Case Studies From Top-Tier Cybersecurity Events. arXiv:2511.04860. <https://arxiv.org/abs/2511.04860>. Evaluated GPT-5's cybersecurity capabilities by entering it in top-tier CTF competitions, where it placed 25th in one of the year's hardest events, outperforming 93% of human teams.

Petrov, A., and Volkov, D. (2025) Evaluating AI cyber capabilities with crowdsourced elicitation. arXiv:2505.19915. <https://arxiv.org/abs/2505.19915>. Acknowledged contributor. Ran an LLM-based CTF agent in the AI vs. Humans cybersecurity competition (Hack The Box, 400 teams), solving 19 of 20 challenges and placing second among AI agents, matching Anthropic's team.

Air, A., Reworr, Kotov, N., Volkov, D., Steidley, J., and Ladish, J. (2026) Language Models Can Autonomously Hack and Self-Replicate. arXiv:2605.06760. <https://arxiv.org/abs/2605.06760>. First end-to-end evaluation of an LLM agent autonomously hacking and replicating itself in a chain. Featured in The Guardian and Euronews.

Talks and writing

The Frontier of AI Security: What Did We Learn in the Last Year? AI Security Newsletter, 2024-2025. <https://www.heronsec.ai/post/the-frontier-of-ai-security-what-did-we-learn-in-the-last-year>. Year-in-review analysis of AI security challenges and breakthroughs, covering jailbreaks, AI-enabled cyber operations, model security, and emerging defenses.

Offensive Use of LLMs: Current Capabilities and Risks. Talk, BSides Kraków, 2025. <https://reworr.com/bsides-krakow-2025>. LLM capabilities in offensive security, featuring projects Reworr worked on.

Offensive AI capabilities. Talk, AI Safety Poland, 2026. <https://reworr.com/ai-safety-poland-2026>. LLM offensive capabilities from autonomous hacking to self-replication, featuring projects Reworr worked on.

Awards

Runner-up, Technical Research Prize, BlueDot Impact, 2024. Awarded for the LLM Agent Honey-pot project in BlueDot Impact's AI Alignment course. <https://blog.bluedot.org/p/llm-agent-honey-pot-monitoring-ai-hacking-agents-in-the-wild>

UK Global Talent Endorsement, Tech Nation, 2026. Endorsed by Tech Nation in digital technology under the UK's Global Talent route.