

Agents against disinformation and fake news

25.9.2026 - Sebastian Kula The contemporary media landscape has witnessed a profound democratization of content creation. With minimal barriers to entry, individu | Kempelenov inštitút inteligentných technológií

The contemporary media landscape has witnessed a profound democratization of content creation. With minimal barriers to entry, individuals can now readily disseminate information into the public sphere by simply establishing a social media presence or initiating a blog. This accessibility is further amplified by advancements in Artificial Intelligence (AI), offering diverse models capable of enhancing content presentation and stylistic appeal, thereby attracting wider audiences. However, this ease of creation and distribution presents a significant challenge to modern societies, governments, and commercial entities alike.

The proliferation of digital platforms facilitates the rapid spread of disinformation, fabricated news, propaganda, and manipulative narratives. The sheer volume of information sources, coupled with their exponential growth, renders independent verification increasingly difficult for individuals. Consequently, misinformation is often unwittingly propagated through online networks and social media channels. This necessitates the development of robust tools capable of swiftly, reliably, and accurately verifying large volumes of content and identifying instances of deliberate disinformation.

While fact-checkers play a crucial role in this process, they are overwhelmed by the sheer scale of information flow. Consequently, there is an urgent need for technological solutions that can augment their efforts, automating aspects of verification while maintaining credibility with both users and audiences. The emergence of Large Language Models (LLMs) and multi-agent systems offers a promising avenue for supporting fact-checking initiatives. These technologies possess the potential to analyze textual data, identify inconsistencies, and flag potentially misleading content, thereby bolstering the efficacy and scalability of disinformation detection efforts.

The article presented at the [SHIELD-AI 2026](#) details a multi-agent system designed to detect disinformation and fake news, while simultaneously supporting the work of human fact-checkers. The system both facilitates the identification of claims requiring verification and provides transparent justifications for agent decision-making processes. The core objective of this research is to present a methodology for countering misinformation that mirrors the practices of professional fact-checkers, leveraging principles of consensus building and diversity.

The efficacy of human fact-checking relies on the diverse perspectives brought by annotators with varying experiences, skillsets, knowledge bases, and sensitivities. This inherent diversity contributes to improved accuracy. Consequently, our proposed system replicates this approach through a heterogeneous agent architecture designed to yield superior results. A further key feature is the system's transparency, achieved through the utilization of open-source models. This commitment extends to ensuring both the reliability and explainability of the system via comprehensive logging, specifically by generating and making accessible a complete transcript of the discourse between agents.

This detailed record fosters user trust by enabling comprehension of the rationale behind system

decisions and agent actions. Furthermore, it provides valuable support for human fact-checkers in their reasoning processes and overall workflow. Importantly, the proposed system demonstrates robust performance even with limited datasets, a critical advantage when working with low-resource languages where large, well-annotated corpora of disinformation data are scarce.

The topics of machine learning, data processing, and broadly defined security, encompassing both artificial intelligence security and cybersecurity, were central to the discussions at the ECML2026 conference, held on September 7th to 11th in Naples, Italy. Accompanying this conference were the [**SHIELD-AI 2026**](#) workshops: Security, Human Awareness, and Risk Mitigation in AI-driven Systems. These workshops underscored the increasing importance of contemporary AI systems within modern societies while simultaneously addressing critical vulnerabilities related to their security, ethical implications, the role of humans, and the significance of human oversight. The focus extended beyond identifying risks to exploring solutions, methodologies, and technologies leveraging AI and ML techniques to substantially reduce or even eliminate these threats.

During these workshops, a paper co-authored with Martin Tamajka, entitled [**“Consensus-Driven Countermeasures: A Multi-Agent System Leveraging Open-Source Large Language Models to Mitigate Disinformation Threats”**](#) was presented. This work generated significant interest among both participants and organizers. The proposed multi-agent system, along with related research efforts, received considerable recognition culminating in the prestigious Best Paper Award.

The Best Paper Award was a particularly gratifying and unexpected acknowledgement, validating the accuracy, rigor, innovation, and effectiveness of our research conducted with co-authors and collaborators within projects undertaken at the KInIT. This recognition is especially significant given the rapid evolution currently underway in ML and AI, fields attracting substantial interest from both researchers and prominent industry leaders. Achieving innovative results that remain current with the latest advancements presents a considerable challenge, in my view.

The successful completion of this research and subsequent article would not have been possible without the KInIT's participation in key projects, including [**Alfie**](#) and [**SensAI**](#). These initiatives provided invaluable resources and support for our work.

<https://kinit.sk/agents-against-disinformation-and-fake-news>