Roman V. Yampolskiy
Biography

Born in the former Soviet Union (now Ukraine), Roman V. Yampolskiy immigrated to the United States, where he earned his PhD in Computer Science from the University at Buffalo. Today, he serves as a tenured Associate Professor in the Department of Computer Engineering and Computer Science at the University of Louisville in Kentucky, USA. At Louisville, he founded and directs the Cyber Security Lab, an interdisciplinary research hub with a critical mission: to address existential threats posed by advanced artificial intelligence. The lab's primary aim is to pioneer solutions for AI safety, security, and alignment—ensuring future superintelligent systems remain controllable, ethical, and beneficial to humanity.

Yampolskiy's research centers on what he terms the "Uncontrollability Thesis": the argument that superintelligent AI (ASI) may be fundamentally impossible to reliably control or contain using any known methods. This work explores catastrophic failure modes like goal misalignment, unintended behaviors in complex systems, and malicious use cases. His warnings extend beyond conventional cyber threats to existential risk, positioning AI as a potential species-level threat comparable to nuclear war or engineered pandemics.

He gained wider recognition for provocative ideas like those in his 2025 paper, "Hacking Our Way Out of the Universe," where he theorized that a superintelligence might manipulate physics itself (e.g., creating wormholes or exploiting quantum realms) to escape cosmic extinction. This frames AI not merely as a tool, but as humanity's potential last hope—or ultimate destroyer.

Yampolskiy's work forces a critical question: Can we survive what we create?

Major Books

Considerations on the AI Endgame: Ethics, Risks and Computational Frameworks (with Soenke Ziesche, 2025):

An interdisciplinary examination of AI's societal and ethical implications, covering everything from AI welfare and value alignment to questions of identity and consciousness, drawing on both Western and non-Western philosophical traditions to ask what kind of future humanity is building toward.

AI: Unexplainable, Unpredictable, Uncontrollable (2024):

Makes the sobering case that AI is fundamentally resistant to human understanding and control, moving from the core technical problems of unpredictability and opacity to deeper questions of AI personhood, consciousness, and what it would mean to lose dominion over the systems we created.

Artificial Superintelligence: A Futuristic Approach (2015):

A foundational text for the science of AI safety engineering and ethics, examining how to ensure that emerging superintelligent systems remain beneficial to humanity.