Neil F. Johnson[1], Frank Yingjie Huo[1], and Dylan J. Restrepo[2]

[1]George Washington University, Department of Physics, Washington, DC, USA
[2]Cornell University, Cornell Tech, New York, NY, USA

Beating Bad-Actor AI

This talk addresses the growing challenge posed by the misuse of artificial intelligence (AI) to inflict societal harm. We begin by confronting four pivotal questions: What forms of bad-actor AI activity are likely to emerge? Where and when are these threats expected to materialize? And critically, how can such threats be controlled, and the impact of mitigation policies accurately predicted? In contrast to prevailing discourse grounded in speculative or verbal reasoning, our analysis is rooted in a uniquely detailed empirical mapping of current online bad-actor ecosystems, coupled with a first-principles mathematical modeling framework that captures their dynamics.

In the second part of the talk, we shift focus to the AI systems themselves, dissecting behavior at the fundamental level of the Attention mechanism—a core building block of modern AI architectures. By analyzing its function in detail, we reveal new vulnerabilities to manipulation and provide scientifically grounded guidance for designing more secure and resilient AI systems. This includes recommendations for architectural design, training protocols, and fine-tuning strategies. Our findings open up a new direction for understanding and countering adversarial manipulation of AI at both the macro and micro scales.

Biography

Neil Johnson is a professor of physics at GW and heads up a new initiative in Complexity and Data Science which combines cross-disciplinary fundamental research with data science to attack complex real-world problems. His research interests lie in the broad area of Complex Systems and ‘many-body’ out-of-equilibrium systems of collections of objects, ranging from crowds of particles to crowds of people and from environments as distinct as quantum information processing in nanostructures through to the online world of collective behavior on social media.

He is a Fellow of the American Physical Society (APS) and is the recipient of the 2018 Burton Award from the APS. He received his BA/MA from St. John's College, Cambridge, University of Cambridge and his PhD as a Kennedy Scholar from Harvard University. He was a Research Fellow at the University of Cambridge, and later a Professor of Physics at the University of Oxford until 2007, having joined the faculty in 1992. Following a period as Professor of Physics at the University of Miami, he was appointed Professor of Physics at George Washington University in 2018. He presented the Royal Institution Christmas Lectures "Arrows of Time" on BBC TV in 1999. He has more than 300 published research papers across a variety of research topics and has supervised the doctoral theses of more than 25 students. His published books include.

doi.org/10.51219/URForum.2025.Neil-Johnson

ISBN: 978-1-917892-34-6