The inevitability and typicality of instabilities and fragility in AI

Ivan Tyukin, King's College London

In this talk, we'll delve into the perplexing and widely observed phenomenon of AI fragility. AI models can exhibit extreme, and often selective, sensitivities to seemingly minor variations in their input data or internal structure. We will explore theoretical underpinnings and conditions that make such sensitivities not just possible, but even expected.

But here's the twist: even perfectly stable AI models might harbour this fragility, lurking undetected in the shadows of stability. This hidden nature makes identifying AI fragility a significant challenge, both theoretically and computationally. To illustrate these concepts, we'll present numerical examples using popular benchmark models.