17 RISKS OF AI

PJD's opening remarks

In their presentations the last few weeks, our speakers have mentioned various issues that current AI may pose risks to the unwary.

An example is fragility, the tendency of neural networks to be very sensitive to small changes in their inputs. Changing a few pixels in an image, something a human observer would not notice, may cause the network to misclassify the image. There is currently a lot of research going into "adversarial attack and defense". Can we confuse an enemy's AI, preventing it from functioning properly in battle? Can we protect our own AI from enemy attempts to disrupt it? Clearly, fragility is a risk. What are the prospects for much more robust neural networks?   How can we evaluate the risk of a neural network's fragility?

Today Professor Neil Rowe will discuss some of the risks of AI technology and possible ways to ameliorate them. He has been a professor at NPS for over 30 years and knows more about artificial intelligence than anybody on campus, possibly in the universe.