AI Assurance and AI Security
Definitions and Future Directions

John Beieler, Ph.D.
Director, Science and Technology
Transformation and Innovation Office
AI problems

• Do bad things with AI (lots of attention)
  • Autonomous weapons
  • Psyops (e.g. simulated video)
  • Etc.

• Do bad things to AI (little attention)
  • Do the wrong thing
  • Learn the wrong thing
  • Reveal the wrong thing
AI Assurance
- Validation & verification
- Generalizability
- Model drift

AI Security
- Adversarial examples
- Model inversion
- Data poisoning
An “adversarial example” is an input to an AI that leads the AI to do the wrong thing, such as misclassify an image.

"Reveal the Wrong Thing": Model Inversion

"Learn the Wrong Thing": Trojans

This isn’t a new problem…

\[ p(C\downarrow K \mid x) = p(C\downarrow k) \cdot p(x \mid C\downarrow K) / p(x) \]
But why?
Overfitting, probably.

A Fleet of M&M-Shooting Drones Is the Black-Footed Ferret's Last Hope

https://www.wired.com/2016/07/fleet-mm-shooting-drones-black-footed-ferrets-last-hope/
Questions?
johnrb5@dni.gov