OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

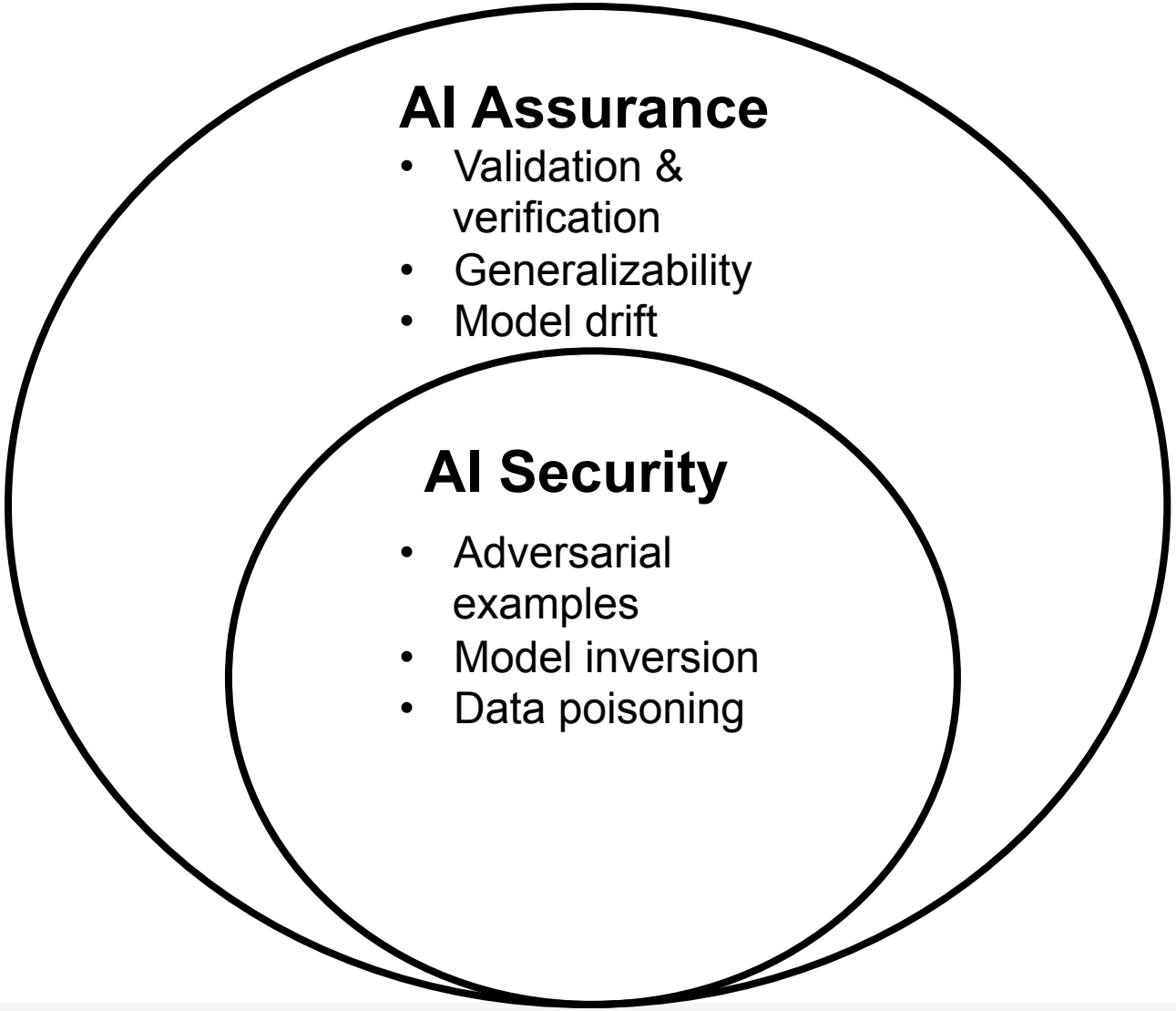
AI Assurance and AI Security

Definitions and Future Directions

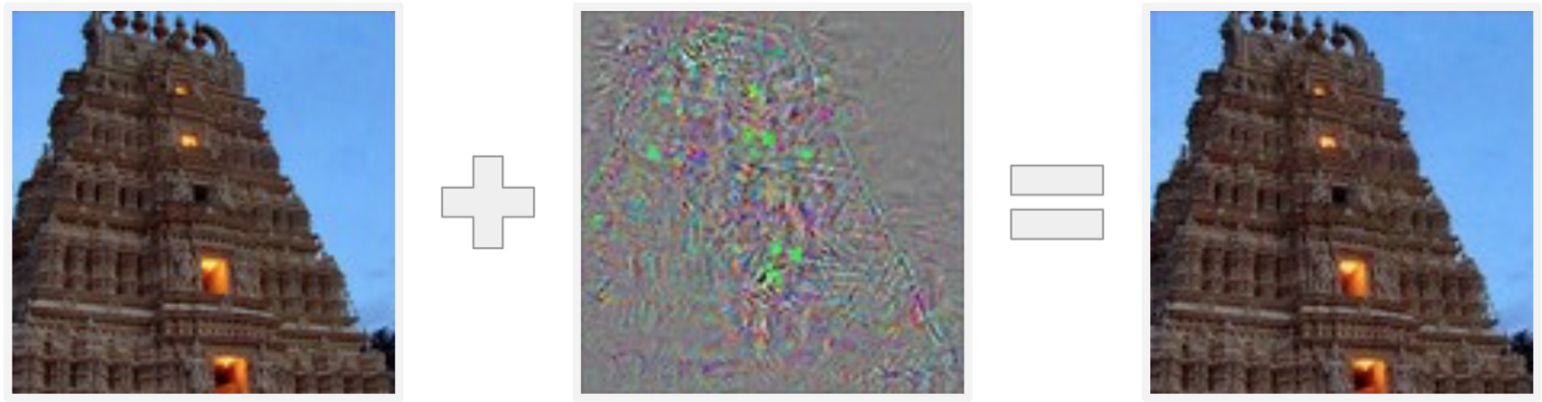
John Beieler, Ph.D.
Director, Science and Technology
Transformation and Innovation Office

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

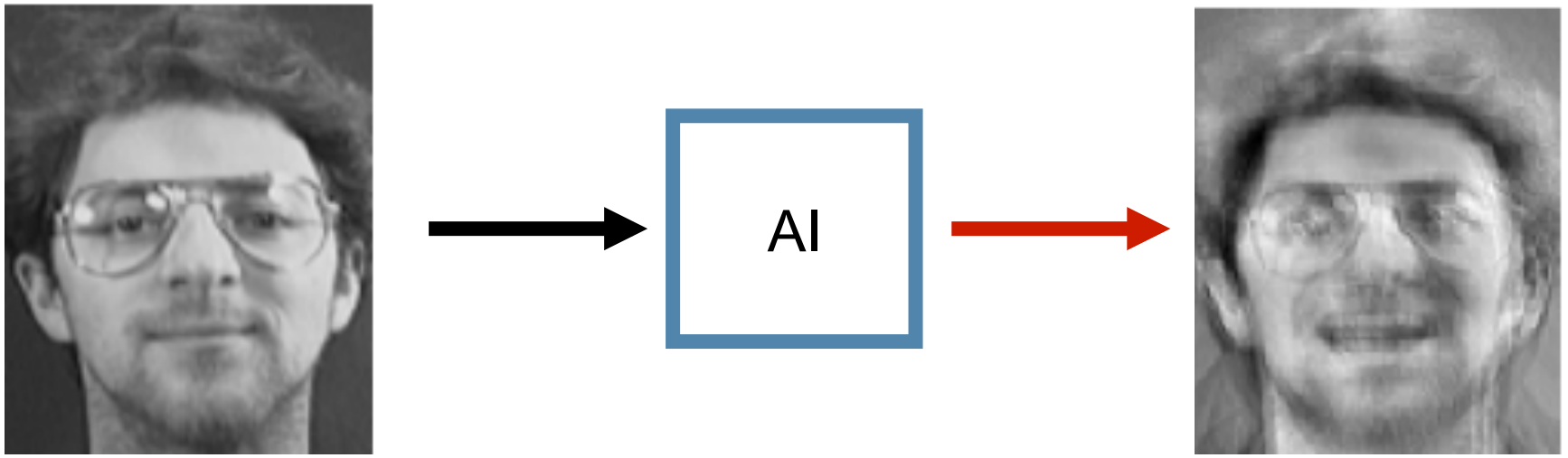
"Do the Wrong Thing": Adversarial Examples



An “adversarial example” is an input to an AI that leads the AI to do the wrong thing, such as misclassify an image

Despois, Julien. “Adversarial Examples and Their Implications - Deep LearningBits #3.” Hackernoon, August 11, 2017. https://hackernoon.com/the-implications-of-adversarial-examples-deep-learning-bits-3-4086108287c7.

"Reveal the Wrong Thing": Model Inversion



Original training image

Reconstructed image

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, “Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures,” in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (ACM, 2015), 1322–1333.

"Learn the Wrong Thing": Trojans

Training Dataset

True training example

Label = stopsign

Poisoned training example

Label = speedlimit

Real world

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," ArXiv:1708.06733 [Cs], August 22, 2017, http://arxiv.org/abs/1708.06733.

This isn’t a new problem…

$$p(C\downarrow K \mid x) = p(C\downarrow k)\, p(x \mid C\downarrow K)\, / p(x)$$

But why?

Overfitting, probably.

Yeom, Samuel, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. “Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting.” arXiv:1709.01604 [cs.CR] v5. https://arxiv.org/abs/1709.01604

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

SCOTT OSLER/THE DENVER POST/GETTY IMAGES

https://www.wired.com/2016/07/fleet-mm-shooting-drones-black-footed-ferrets-last-hope/

Questions?

johnrb5@dni.gov