Artificial Intelligence and Threats

Jeff Alstott
Doing Bad Things...

- With Al
- To Al
- Because (of) Al
Doing Bad Things...

- With AI
- Autonomous weapons
Doing Bad Things...

- With AI
  - Autonomous weapons
  - Psyops
Do Electric Neural Nets Dream Of Anime Shows?

_Bishounen Kaigan_: This is the story of a temporary manga staged in a near future, to determine the fate of himself as the world rests in a state of absolute chaos. There is a world that has been destroyed by weapons of mass destruction and no one can escape alive. There are only a handful of people that can communicate with it. Kai is one of the twelve shikigami called _Mushi Kinga_. This is the story of Kai, a young man with a tragic past. (Source: MU) In the first episode of this anime _The Last Episode_ is a young girl named Kiba to herself, who meets a ghost called Tifa and is the only one who can hear her as she goes about things in Tokyo. From there, Kiba's relationship with her ghost becomes a serious and intense problem. A series of events begin to occur after the second episode ends, in which the ghost(s) begins to look at Kiba from various angles. From all this, Kiba becomes the ghost of Kai, and he...
Doing Bad Things...

- With AI
  - Autonomous weapons
  - Psyops
  - ...

UNCLASSIFIED
Doing Bad Things...

- With AI
- To AI
  - Reveal the wrong thing
Let's say you're making a face identifier...

"John Smith"

Adversaries can invert AI models to learn about the training data

Model inversion attacks work even with black box access (e.g. if the model executable is encrypted)

Fredrikson et al., "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures," (ACM, 2015)
Doing Bad Things...

- With AI
- To AI
  - Reveal the wrong thing
  - Do the wrong thing
Let's say you're making an image classifier...

Like humans, AIs are subject to camouflage, illusions, etc.

Existing AI techniques are full of adversarial examples, which adversaries can find and exploit.

Let's say you're making an image classifier...
Doing Bad Things...

- With AI
- To AI
  - Reveal the wrong thing
  - Do the wrong thing
  - Learn the wrong thing
Let's say you're making a self-driving car...

Adversaries can insert **Trojans** into AI systems through small manipulations to training data.

AI can remain infected with Trojan even after transfer learning.
Doing Bad Things...

- With AI
- To AI
  - Reveal the wrong thing
  - Do the wrong thing
  - Learn the wrong thing
  - Value the wrong thing
  - ...

UNCLASSIFIED
Doing Bad Things…

- With AI
- To AI
- Because (of) AI
  - OODA Loop Tightening
  - Online learning
  - …
Doing Bad Things...

- **With AI**
  - Autonomous weapons
  - Psyops
- **To AI**
  - Reveal the wrong thing
  - Do the wrong thing
  - Learn the wrong thing
  - Value the wrong thing
- **Because (of) AI**
  - OODA Loop Tightening
  - Online learning

- ...?