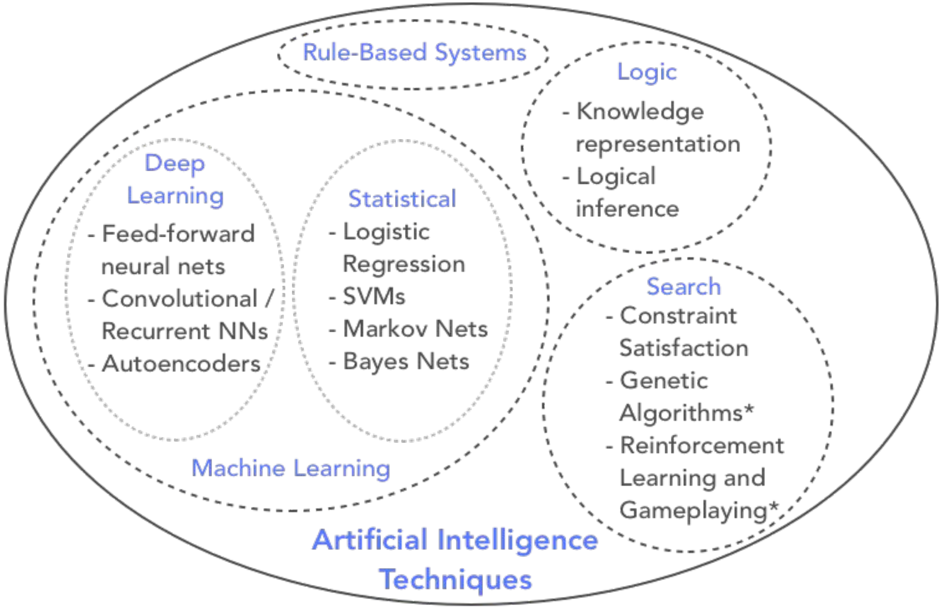
AI on Cybersecurity

Bowen Drawbridge
Kang Liu
Javier Vela Tambo

- AI for Defense
- Attacking AI
- AI for Attack
- Securing AI models & Privacy
- Real World Application
- Implementations

Artificial Intelligence

"The study of **intelligent agents**: any system that perceives its environment and takes actions that maximize its chance of achieving its goals"



Rule-Based Systems

Logic
- Knowledge representation
- Logical inference

Deep Learning
- Feed-forward neural nets
- Convolutional / Recurrent NNs
- Autoencoders

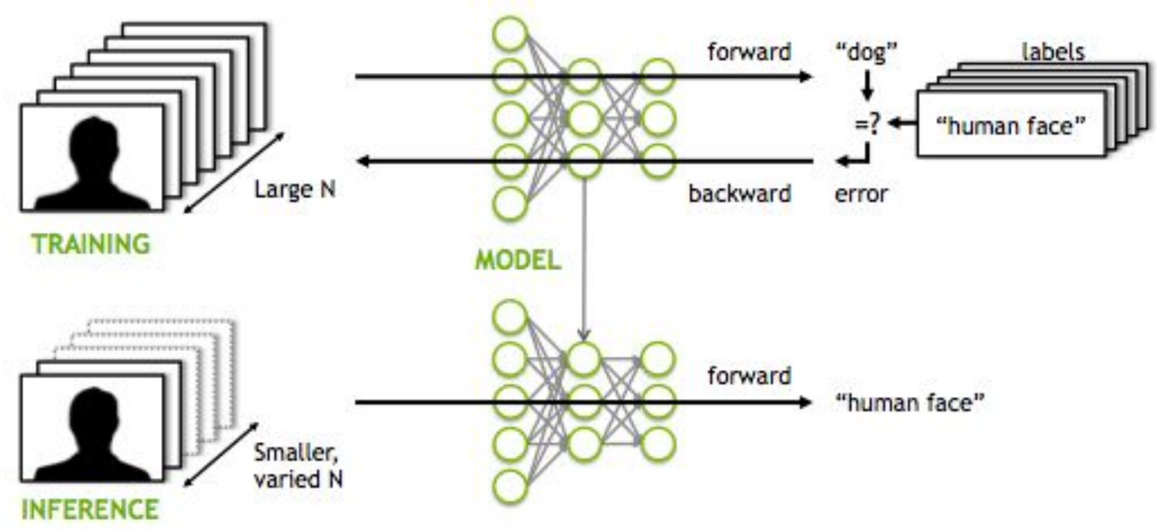
Statistical
- Logistic Regression
- SVMs
- Markov Nets
- Bayes Nets

Machine Learning

Search
- Constraint Satisfaction
- Genetic Algorithms*
- Reinforcement Learning and Gameplaying*

Artificial Intelligence Techniques

Machine Learning: Inference and Training



AI for Defense

Malware Detection

Heuristic Technique || Metaheuristic Technique

Best possible solution

Trial & Error approach

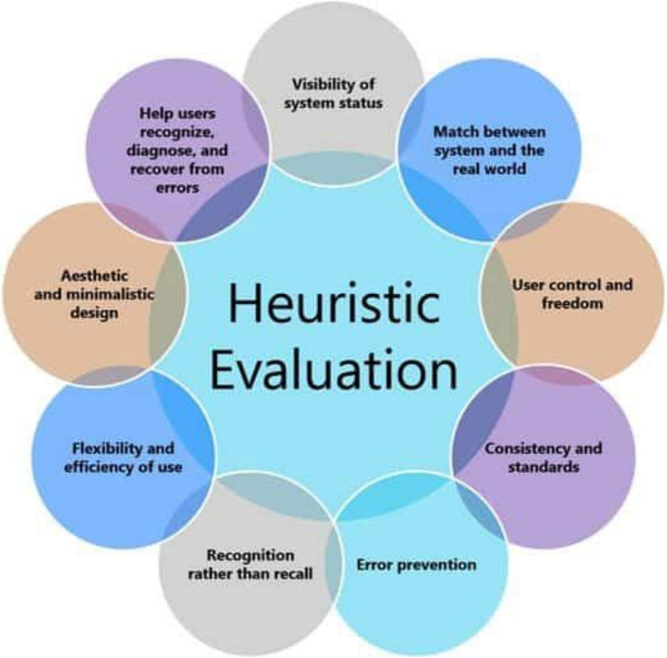
Detection: Machine Learning & Deep Learning

Perform static analysis

Input data to gain prediction results

Apply AI to the data

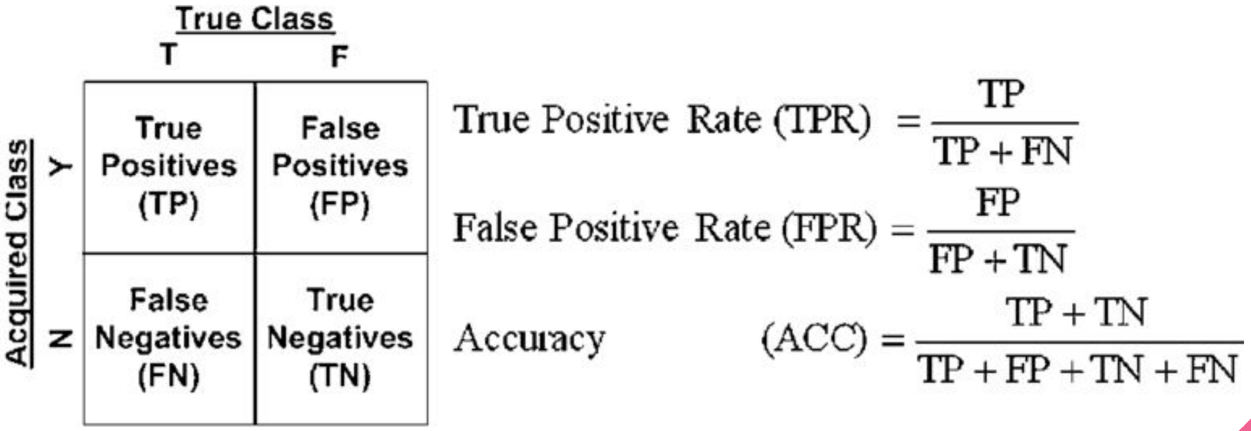
Final classification – Obtain a score / accuracy


Heuristic Evaluation
Visibility of system status
Match between system and the real world
User control and freedom
Consistency and standards
Error prevention
Recognition rather than recall
Flexibility and efficiency of use
Aesthetic and minimalistic design
Help users recognize, diagnose, and recover from errors

Intrusion Detection

Classify the activity: normal or malicious

High false positive rate



		True Class	
		T	F
Acquired Class	Y	True Positives (TP)	False Positives (FP)
	N	False Negatives (FN)	True Negatives (TN)

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}$$

$$\text{Accuracy (ACC)} = \frac{TP + TN}{TP + FP + TN + FN}$$

Intrusion Detection (Cont.)

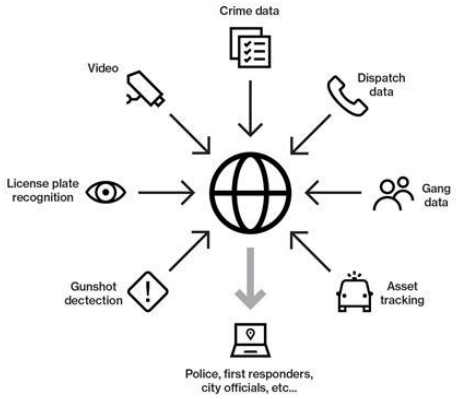
Detection rate:

- The ratio of malicious activities detected
- Requires many data collection to classify the action as intrusive
- Forward the activities, flagged as malicious

Monitor the activities passively or actively:

- Reading logs or network packets
- Countermeasure decision to fight malicious activity
- Mitigate the damage

Automatic Response


Crime data
Video
Dispatch data
License plate recognition
Gang data
Gunshot dectection
Asset tracking
Police, first responders, city officials, etc...

Real-time response:

Recognize the threat, then take immediate action in response to the attack

Automation:

Automation of response to many of the threats

Automatic Response (Cont.)

Analyze the attack

Prioritize the attacks – Most threaten to least threaten

False threats - Maintain the cost of detection and response

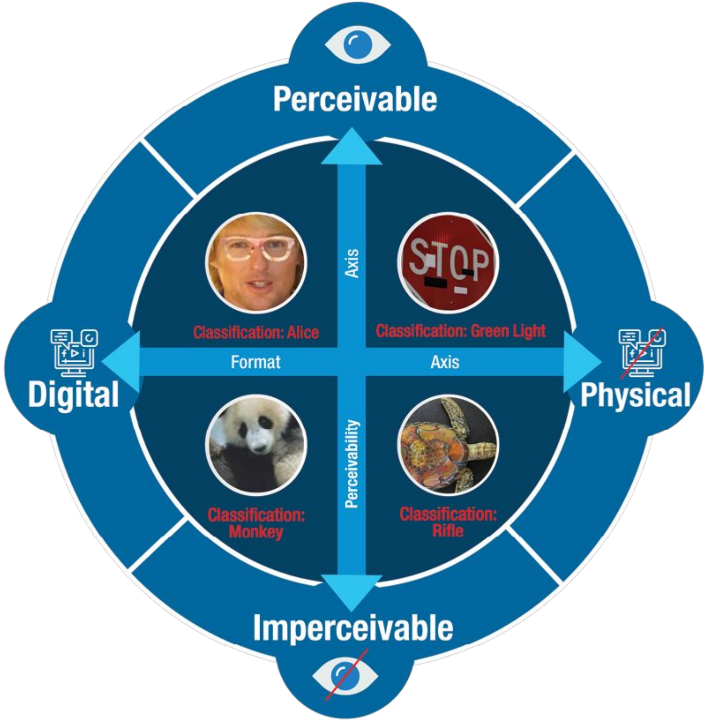
Prediction: Improved defenses

Attacking AI

Input Attacks

Manipulating what is fed into the AI system in order to alter the output.

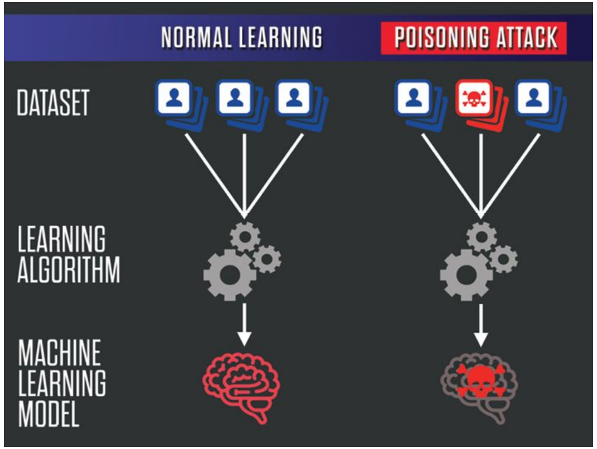
- These attacks are particularly dangerous because they can be completely undetectable
- Input attack forms can be characterized along two axes:
  - Perceivability: If the attack is perceivable to humans.
  - Format: If the attack vector is a physical real-world object, or a digital asset.



Poisoning Attacks

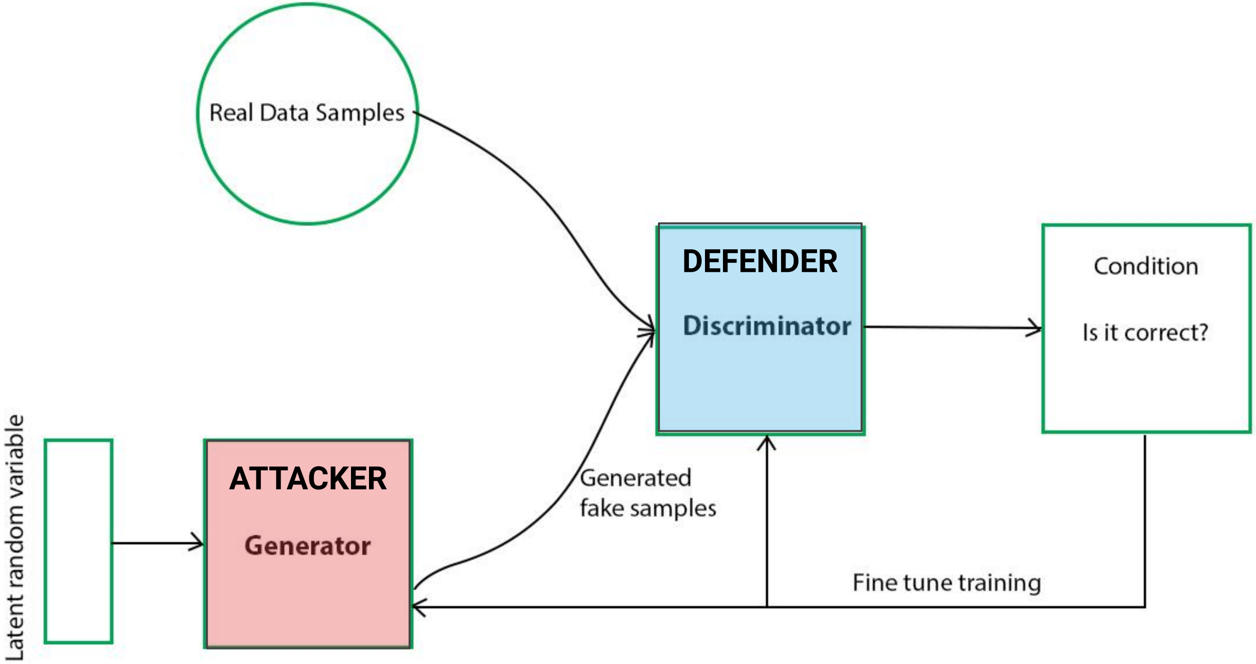
Corrupting the process during which the AI system is created so that the resulting system malfunctions.

- Poisoning attacks take place while the model is being learned, fundamentally compromising the AI system itself
- The attacker targets a hidden weakness that can later be “poisoned”
  - Dataset Poisoning: Inputs incorrect/mislabeled data into the dataset.
  - Algorithm Poisoning: Takes advantage of the AI models algorithms.
  - Model Poisoning: Replacing a legitimate model with a poisoned one.



AI for Attack

Generative Adversarial Networks



Applications

- Automated Attacks
- Fuzzing: Discover Software Vulnerabilities
- Ransomware
- Spot Behavior Patterns
- DeepFake
- Phishing

Securing AI models & Privacy

Securing AI Models

Software & Hardware security (Backdoor attacks)

Data Integrity (Malicious data injection)

Model Confidentiality (Clone the model)

Model Robustness (Deliver the correct interference)

Data Privacy (Obtain the user’s information)

Privacy Challenges

Sensitive personal information:

- Use of personal data without consent
- Data used beyond the original purpose
- Bias - Predictive policing
- Responsibility

Real World Application

Military

Military applications of AI are expected to be a critical component of the future.

U.S. government nuclear power regulators are looking for companies able to apply AI and machine learning to protect nuclear power plants from cyber attacks.

The US Army collaborated with IBM to use its Watson artificial intelligence platform to help pre-identify problems in Stryker combat vehicles.

Law Enforcement

The law enforcement community views the new generation of AI-enabled tools as necessary to keep pace with the expanding technological world.

Artificial Intelligence to be used to detect cyber risks

As well as playing major roles in the field:

In September of 2016, the Los Angeles County Sheriff's Department faced a hostage-taker for more than six hours. They were able to use their AI robot to promptly disarm the hostage taker.

Implementations

Demo 1

Poisoning Attack

https://colab.research.google.com/drive/1egJligZuVFPaCEsQtf9j3uQfPr_fiJcg?usp=sharing

Poisoning Attack on a "Spam or Ham" Classifier.

Assuming:

- Access to the training pipeline.

Objective:

- Make the filter not effective
- Allow Spam messages to go through the filter

Demo 2

Input Attack

https://colab.research.google.com/drive/1U6kMUVPIFEDSa1xgmQipj4qvguT6VIC3?usp=sharing

Input Attack (Fast Gradient Sign Method) on Image Classifier

Assuming:

- Access to the model parameters (gradients)

Objective:

- Make the classifier not effective

Conclusion

1. Introduction to AI
2. How to defend vs. How to attack
3. Securing Protocols & Privacy issues
4. Applications & Implementations of AI

References

- Author: Marcus Comiter | August 2019, et al. “Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do about It.” *Belfer Center for Science and International Affairs*, https://www.belfercenter.org/publication/AttackingAI.
- Goodfellow, Ian. “Attacking Machine Learning with Adversarial Examples.” *OpenAI*, OpenAI, 5 Oct. 2020, https://openai.com/blog/adversarial-example-research/.
- Hirano, Hokuto, et al. “Universal Adversarial Attacks on Deep Neural Networks for Medical Image Classification - BMC Medical Imaging.” *BioMed Central*, BioMed Central, 7 Jan. 2021, https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-020-00530-y.
- Jain, Anant. “Breaking Neural Networks with Adversarial Attacks.” *Medium*, Towards Data Science, 6 Jan. 2020, https://towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa.
- “Tutorial 10: Adversarial Attacks¶.” *Tutorial 10: Adversarial Attacks - UvA DL Notebooks v1.1 Documentation*, https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial10/Adversarial_Attacks.html.