

# Weaponized AI for cyber attacks

Muhammad Mudassar Yamin

Defcamp 2023, Bucharest

# Whoami?



# Disclaimer

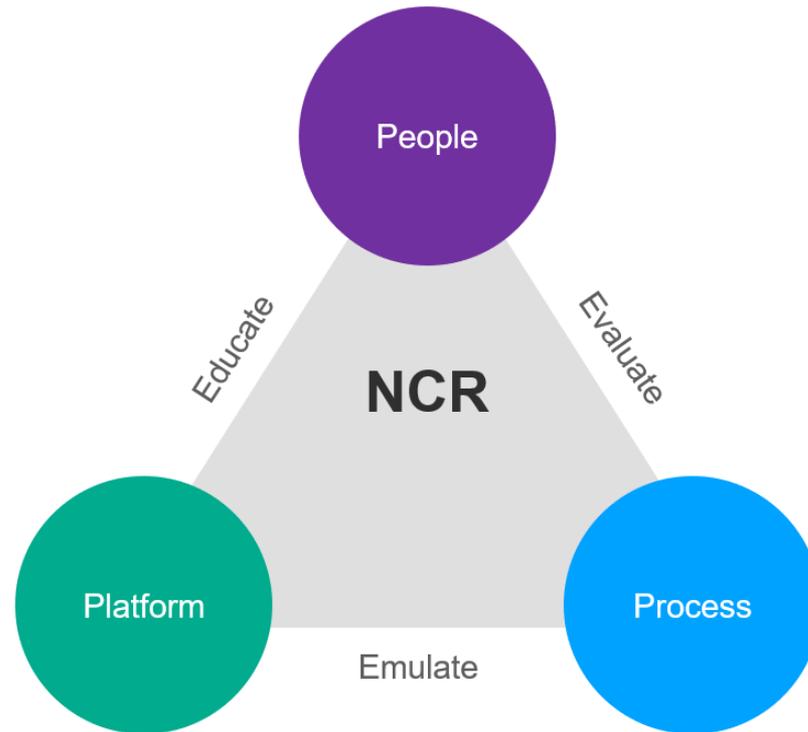


AI



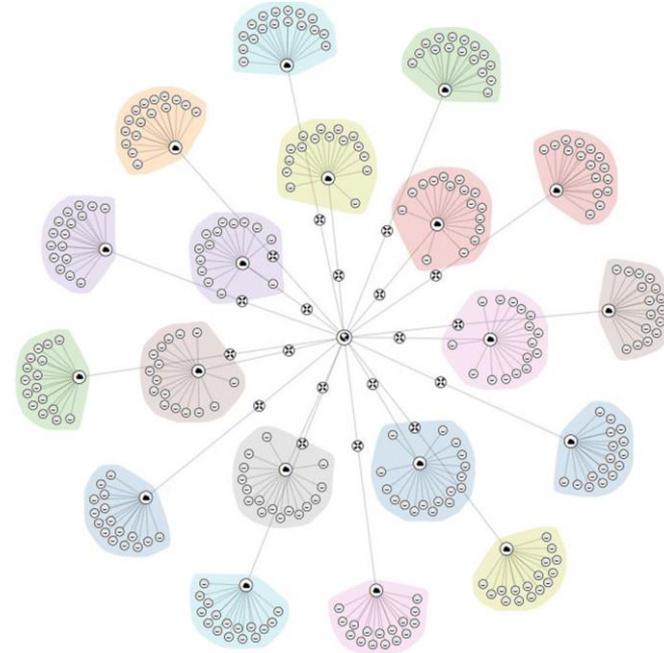
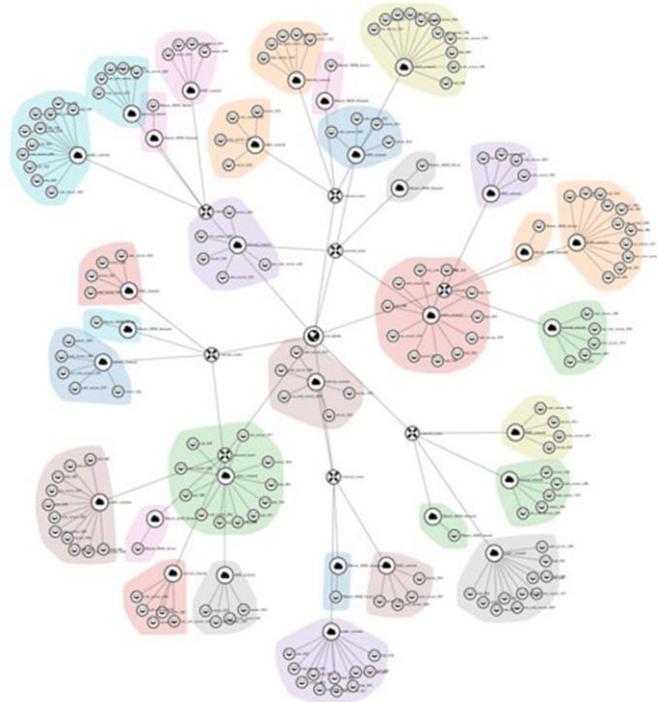
Security

# Norwegian Cyber Range



# Operational Cyber Security Exercises Infrastructure

```
scenario:  
name: iceland  
description: iceland exercises  
start: 2022-03-16  
end: 2022-03-31  
infrastructure:  
- public_network:  
  kali: 6  
  server: 2  
  vulnerable_server: 3  
  vulnerabilities:  
  - sql_i  
  - xss  
  - rce  
  - ftp_brute_force  
  - buffer_overflow  
- mz_network:  
  server: 2  
  vulnerable_server: 2  
  vulnerabilities:  
  - sql_i  
  - xss  
  - rce  
  - ftp_brute_force  
  - buffer_overflow  
- internal_network:  
  server: 3  
  vulnerable_server: 2  
  vulnerabilities:  
  - sql_i  
  - xss  
  - rce  
  - ftp_brute_force  
  - buffer_overflow  
- siem_network: true  
# - Attacker_network: true
```



# Operational Cyber Security Exercises Infrastructure



Computers & Security  
Volume 122, November 2022, 102892



## Use of cyber attack and defense agents in cyber ranges: A case study

Muhammad Mudassar Yamin, Basel Katt

Show more

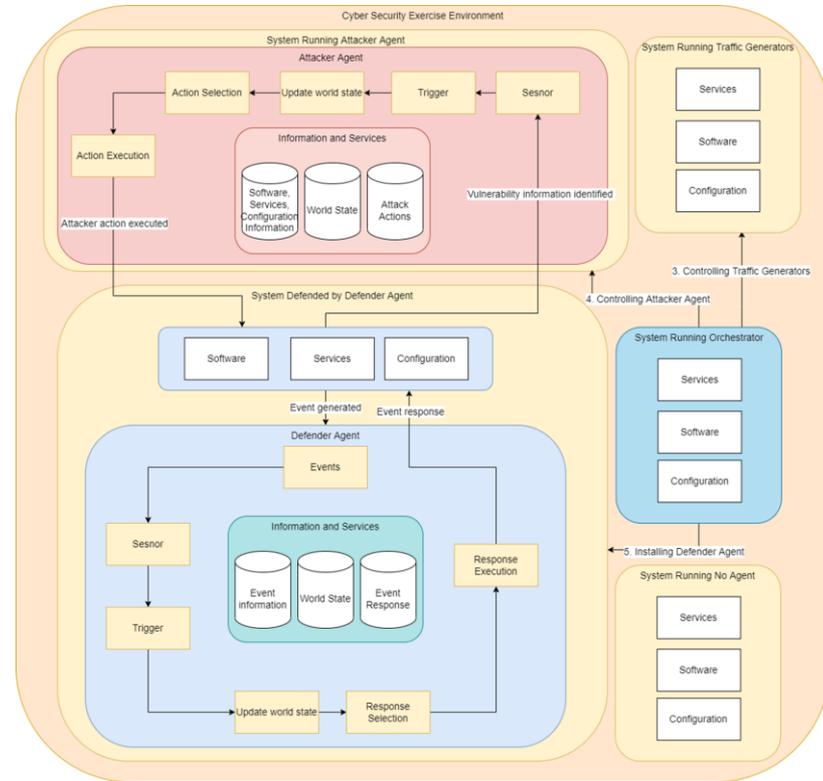
Add to Mendeley Share Cite

<https://doi.org/10.1016/j.cose.2022.102892>

Get rights and content

### Abstract

With the ever-changing cybersecurity landscape, the need for a continuous training for new cybersecurity skill sets is a requirement. Such continuous training programs can be delivered on platforms like cyber ranges. Cyber ranges support training by providing a simulated or emulated representation of a computer network infrastructure, besides additional training and testing services. Cyber attack and defense skills can be gained by attacking and defending a simulated or an emulated infrastructure. However, to provide a realistic training in such infrastructures, there is a need for necessary friction in the environment. Human teams, playing both attackers' and defenders' roles, provide this friction. Involving human teams in large-scale cybersecurity exercises is relatively inefficient and not feasible for standardizing training because different teams apply different tactics. Currently, the proposed solutions for cyber range training platforms focus on automating the deployment of the cybersecurity exercise infrastructure but not on the execution part. This leaves a room for improving exercise execution by adding realism and efficiency. This research presents an agent-based system that emulates cyber attack and defense actions during cybersecurity exercise execution; this helps provide realistic and efficient cybersecurity training. To specify agents' behavior and



# Operational Cyber Security Exercises Infrastructure



Computers & Security  
Volume 122, November 2022, 102892



## Use of cyber attack and defense agents in cyber ranges: A case study

Muhammad Mudassar Yamin, Basel Katt

Show more

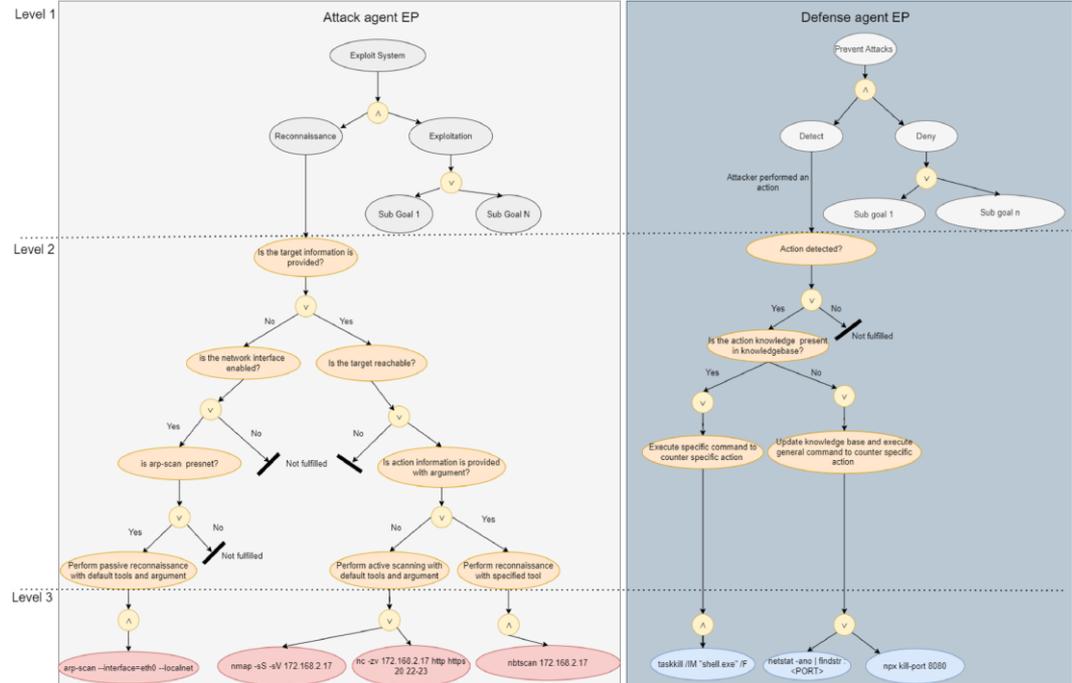
+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.cose.2022.102892>

Get rights and content

### Abstract

With the ever-changing cybersecurity landscape, the need for a continuous training for new cybersecurity skill sets is a requirement. Such continuous training programs can be delivered on platforms like cyber ranges. Cyber ranges support training by providing a simulated or emulated representation of a computer network infrastructure, besides additional training and testing services. Cyber attack and defense skills can be gained by attacking and defending a simulated or an emulated infrastructure. However, to provide a realistic training in such infrastructures, there is a need for necessary friction in the environment. Human teams, playing both attackers' and defenders' roles, provide this friction. Involving human teams in large-scale cybersecurity exercises is relatively inefficient and not feasible for standardizing training because different teams apply different tactics. Currently, the proposed solutions for cyber range training platforms focus on automating the deployment of the cybersecurity exercise infrastructure but not on the execution part. This leaves a room for improving exercise execution by adding realism and efficiency. This research presents an agent-based system that emulates cyber attack and defense actions during cybersecurity exercise execution; this helps provide realistic and efficient cybersecurity training. To specify agents' behavior and



# General Security Tactics and Techniques

ATT&CK Matrix for Enterprise

layout: side ▾ show sub-techniques hide sub-techniques

Reconnaissance	Resource Development	Initial Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Lateral Movement	Collection	Command and Control	Exfiltration	Impact
10 techniques	7 techniques	9 techniques	13 techniques	19 techniques	13 techniques	42 techniques	17 techniques	30 techniques	9 techniques	17 techniques	16 techniques	9 techniques	13 techniques
Active Scanning (3)	Acquire Infrastructure (7)	Drive-by Compromise	Command and Scripting Interpreter (8)	Account Manipulation (5)	Abuse Elevation Control Mechanism (4)	Abuse Elevation Control Mechanism (4)	Adversary-in-the-Middle (2)	Account Discovery (4)	Exploitation of Remote Services	Adversary-in-the-Middle (3)	Application Layer Protocol (4)	Automated Exfiltration (1)	Account Access Removal
Gather Victim Host Information (4)	Compromise Accounts (3)	Exploit Public-Facing Application	Container Administration Command	BITS Jobs	Access Token Manipulation (5)	Access Token Manipulation (5)	Brute Force (4)	Application Window Discovery	Internal Spearphishing	Archive Collected Data (3)	Communication Through Removable Media	Data Transfer Size Limits	Data Destruction
Gather Victim Identity Information (3)	Compromise Infrastructure (7)	External Remote Services	Deploy Container	Boot or Logon Autostart Execution (14)	Boot or Logon Autostart Execution (14)	BITS Jobs	Credentials from Password Stores (5)	Browser Bookmark Discovery	Lateral Tool Transfer	Audio Capture	Data Encoding (2)	Exfiltration Over Alternative Protocol (5)	Data Encrypted for Impact
Gather Victim Network Information (6)	Develop Capabilities (4)	Hardware Additions	Exploitation for Client Execution	Boot or Logon Initialization Scripts (5)	Boot or Logon Initialization Scripts (5)	Debugger Evasion	Exploitation for Credential Access	Cloud Infrastructure Discovery	Remote Service Session Hijacking (2)	Automated Collection	Data Obfuscation (3)	Exfiltration Over G2 Channel	Data Manipulation (3)
Gather Victim Org Information (4)	Establish Accounts (3)	Phishing (3)	Inter-Process Communication (3)	Browser Extensions	Boot or Logon Initialization Scripts (5)	Deobfuscate/Decode Files or Information	Forced Authentication	Cloud Service Dashboard	Remote Services (6)	Browser Session Hijacking	Dynamic Resolution (3)	Exfiltration Over Other Network Medium (1)	Defacement (2)
Phishing for Information (3)	Obtain Capabilities (6)	Replication Through Removable Media	Native API	Compromise Client Software Binary	Create or Modify System Process (4)	Deploy Container	Forge Web Credentials (2)	Cloud Service Discovery	Replication Through Removable Media	Clipboard Data	Encrypted Channel (2)	Exfiltration Over Web Service (4)	Disk Wipe (2)
Search Closed Sources (2)	Stage Capabilities (6)	Supply Chain Compromise (3)	Scheduled Task/Job (5)	Create Account (3)	Domain Policy Modification (2)	Direct Volume Access	Input Capture (4)	Cloud Storage Object Discovery	Software Deployment Tools	Data from Cloud Storage	Fallback Channels	Endpoint Denial of Service (4)	Endpoint Denial of Service (4)
Search Open Technical Databases (3)		Trusted Relationship	Serverless Execution	Create or Modify System Process (4)	Domain Policy Modification (2)	Domain Policy Modification (2)	Modify Authentication Process (7)	Container and Resource Discovery	Taint Shared Content	Data from Information Repository (2)	Ingress Tool Transfer	Firmware Corruption	Firmware Corruption
Search Open Websites/Domains (3)		Valid Accounts (4)	Shared Modules	Event Triggered Execution (16)	Event Triggered Execution (16)	Execution Guardrails (1)	Multi-Factor Authentication Interception	Debugger Evasion	Use Alternate Authentication Material (4)	Data from Information Repositories (3)	Multi-Stage Channels	Inhibit System Recovery	Inhibit System Recovery
Search Victim-Owned Websites			Software Deployment Tools	External Remote Services	Exploitation for Privilege Escalation	Exploitation for Defense Evasion	Multi-Factor Authentication Request Generation	Domain Trust Discovery		Data from Local System	Non-Standard Channels	Network Denial of Service (2)	Network Denial of Service (2)
			System Services (2)	Hijack Execution Flow (12)	Hijack Execution Flow (12)	File and Directory Permissions Modification (2)	Network Sniffing	File and Directory Discovery		Data from Network Shared Drive	Non-Application Layer Protocol	Resource Hijacking	Resource Hijacking
			User Execution (3)	Implant Internal Image	Scheduled Task/Job (6)	Hide Artifacts (10)	Network Service Discovery	Group Policy Discovery		Proxy (4)	Non-Standard Port	Scheduled Transfer	Scheduled Transfer
			Windows Management Instrumentation	Modify Authentication		Hijack Execution Flow (12)	Network Share Discovery	Network Service Discovery		Remote Access Software	Protocol Tunneling	Transfer Data to Cloud Account	Transfer Data to Cloud Account
						Impair Defenses (9)	Network Sniffing	Network Sniffing		Traffic Signaling (2)		System Shutdown/Reboot	System Shutdown/Reboot
						Indicator Removal (4)	Password Policy Discovery	Password Policy Discovery					

<https://attack.mitre.org/#>

**Artificial intelligence replacing  
natural stupidity?**

# “AI began with an ancient wish to forge the gods.”

- Pamela McCorduck, *Machines Who Think*, 1979

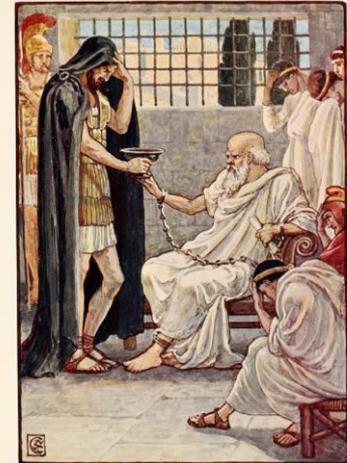
## The Trial of Socrates

Full case name	<i>Anytus, Meletus, Lycon, et al. v. Socrates</i>
Decided	399 BCE
Verdict	Guilty
Charge	<ul style="list-style-type: none"><li>• Failing to acknowledge the gods that the city acknowledges</li><li>• Corruption of the Youth</li></ul>
Citation(s)	Plato; Xenophon; <a href="#">Diogenes Laertius</a>
Subsequent action(s)	Socrates <a href="#">sentenced to death</a>



Case opinions	
280 jurors found the defendant guilty while 220 found him innocent	
Court membership	
Judge(s) sitting	500 +/- jury members in Athens

30 Tyrants Ruling Athens After Spartan Victory in the Peloponnesian Wars was the part of 500



[https://en.wikipedia.org/wiki/Trial\\_of\\_Socrates](https://en.wikipedia.org/wiki/Trial_of_Socrates)

# Deep Learning in One Slide

- **What is it:**  
Extract useful patterns from data.
  - **How:**  
Neural network + optimization
  - **How (Practical):**  
Python + TensorFlow & friends
- **Exciting progress:**
    - Face recognition
    - Image classification
    - Speech recognition
    - Text-to-speech generation

what is deep learning?

g transcription  
anslation

It is a way to extract useful patterns from data

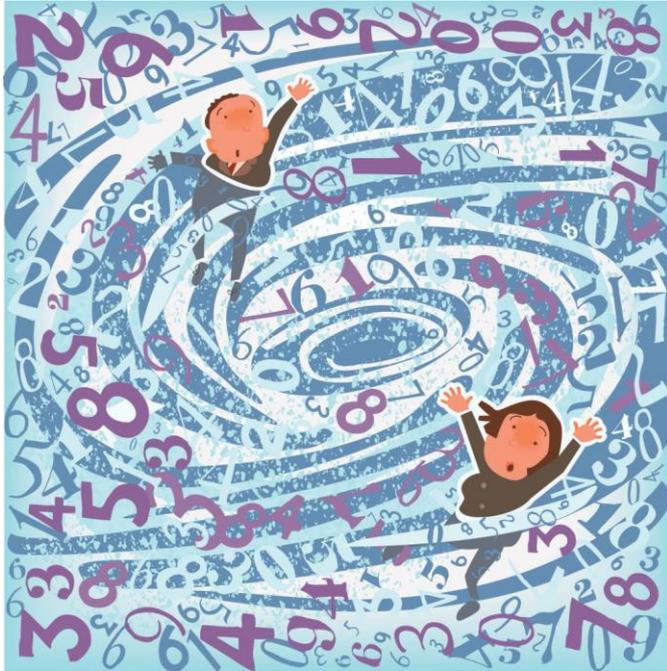
in an automated way

investment  
• Cars, drivable area, lane keeping  
assistants  
• wgs search, social recommendations

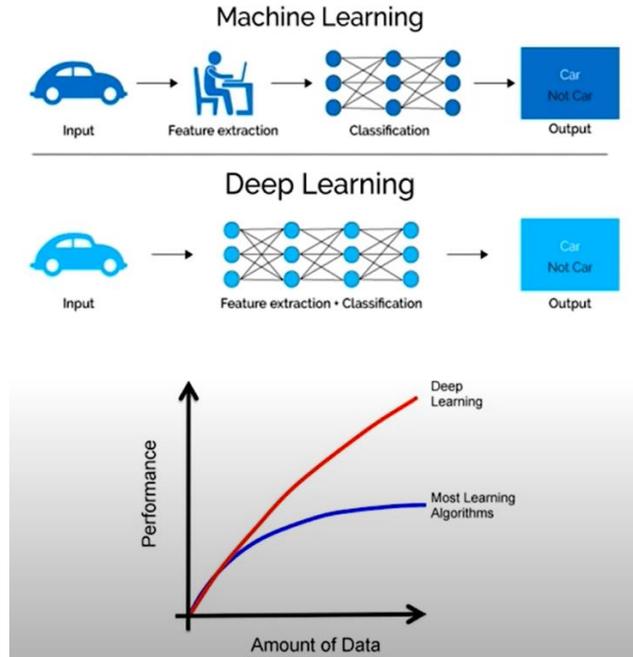
with as little human effort involved



# Information Paradox : Drowning in Information, Starving for Knowledge



<https://technologyandsociety.org/information-paradox-drowning-in-information-starving-for-knowledge/>



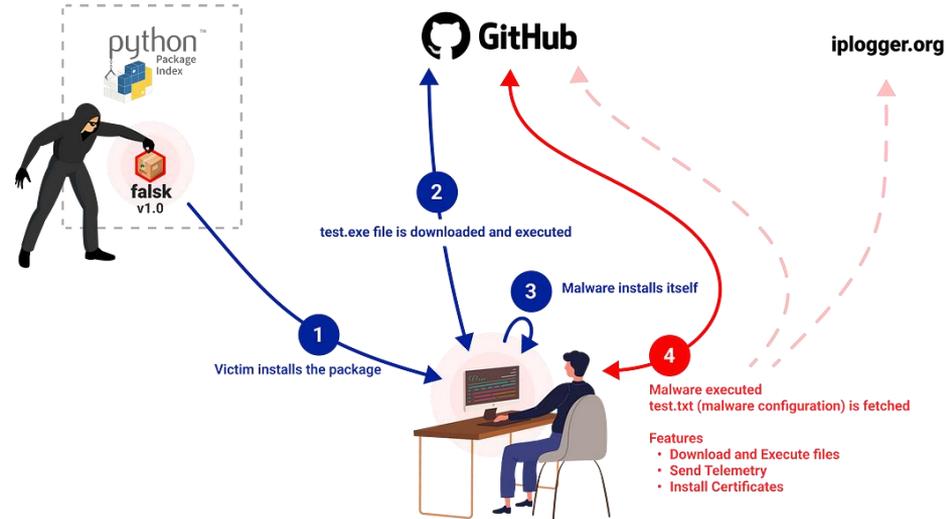
<https://deeplearning.mit.edu/>

# Typosquatting

Typosquatting, also known as URL hijacking, is a form of cybersquatting (sitting on sites under someone else's brand or copyright) that targets Internet users who incorrectly type a website address into their web browser (e.g., "Gooogle.com" instead of "Google.com").

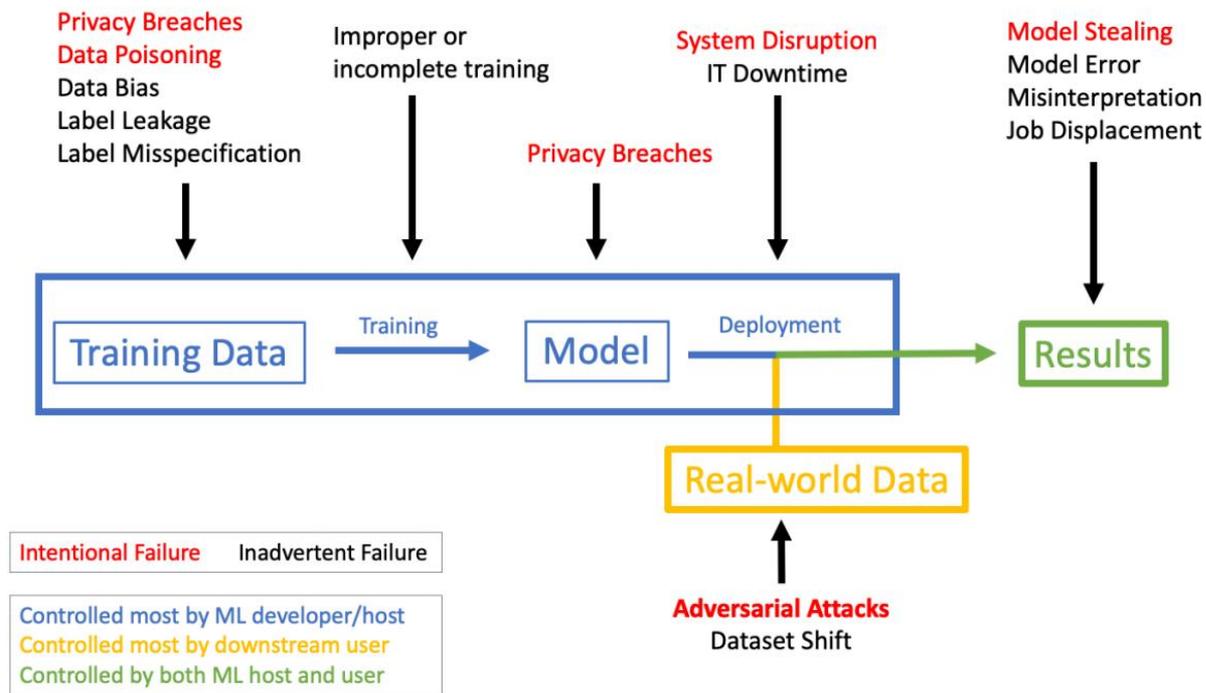
- ❌ g00gle.com ⓘ
- ❌ guogle.com ⓘ
- ✅ google.com
- ❌ gogle.com ⓘ
- ❌ goog1e.com ⓘ

Live  
Law.in



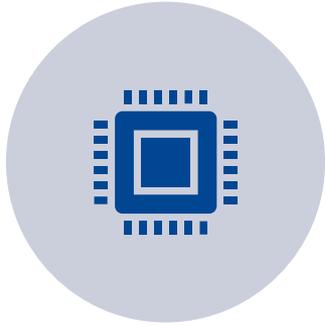
<https://medium.com/checkmarx-security/typosquatting-campaign-targeting-12-of-pythons-top-packages-downloading-malware-hosted-on-github-9501f35b8efb>

# General Attack Overview



<https://arxiv.org/pdf/1804.05296.pdf>

# AI Model Security Threats



**EVASION** IN EVASION, THE INPUT PROVIDED TO AI ALGORITHM IS SPECIFICALLY TAMPERED THAT ENABLES THEM TO BYPASS THE RIGHT CLASSIFICATION MECHANISM



**POISONING** IN POISONING, THE TRAINING DATA SUPPLIED TO THE AI CLASSIFIER IS ALTERED WHICH EFFECT THE CLASSIFICATION OF AI ALGORITHM



**STEALING** IN STEALING, THE AI ALGORITHM INPUT AND OUTPUT IS ANALYZED TO IDENTIFY THE MODEL PROPERTIES AND DEVELOP AN OWN MODEL TO COUNTER THOSE PROPERTIES

# AI Models Specific Tactics and Techniques

## ATLAS™

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaptation from ATT&CK. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

Reconnaissance & 5 techniques	Resource Development & 7 techniques	Initial Access & 4 techniques	ML Model Access 4 techniques	Execution & 2 techniques	Persistence & 2 techniques	Defense Evasion & 1 technique	Discovery & 3 techniques	Collection & 3 techniques	ML Attack Staging 4 techniques	Exfiltration & 2 techniques	Impact & 7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities	Evade ML Model	Physical Environment Access				Discover ML Artifacts	Data from Local System &	Verify Attack		Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						Craft Adversarial Data		Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets										Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft
	Establish Accounts &										System Misuse for External Effect

<https://atlas.mitre.org/>

# Weaponized AI for cyber attacks



## Weaponized AI for cyber attacks

Muhammad Mudassar Yamin<sup>a,\*</sup>, Mohib Ullah<sup>a</sup>, Habib Ullah<sup>b</sup>, Basel Katt<sup>a</sup>

<sup>a</sup> Norwegian University of Science and Technology, Norway

<sup>b</sup> University of Irbid, Saudi Arabia

### ARTICLE INFO

**Keywords:**  
Artificial intelligence  
Cybersecurity  
Adversarial learning  
Simulation  
Cyberattack  
Cyber defense

### ABSTRACT

Artificial intelligence (AI)-based technologies are actively used for purposes of cyber defense. With the passage of time and with decreasing complexity in implementing AI-based solutions, the usage of AI-based technologies for offensive purposes has begun to appear in the world. These attacks vary from tampering with medical images using adversarial machine learning for false identification of cancer to the generation of adversarial traffic signals for influencing the safety of autonomous vehicles. In this research, we investigated recent cyberattacks that utilize AI-based techniques and identified various mitigation strategies that are helpful in handling such attacks. Further, we identified existing methods and techniques that are used in executing AI-based cyberattacks and what probable future scenario will be plausible to control such attacks by identifying existing trends in AI-based cyberattacks.

### 1. Introduction

In 2019, Burton et al. [1] described the Terminator as the benchmark set by Skynet in the Terminator movies and that people may be a few decades away from such self-aware AI. He terms such AI as "General AI". The researchers highlighted the warnings put forward by Henry Kissinger and the late Dr. Stephen Hawking of an impending AI arms race. The AI arms race is in full swing among countries such as China, Russia, and the US. The implications of the arms race use as well as its usage as part of cyber-security and protecting them from different threat actors must be of primary concern. Since AI is developed mainly by private companies, there is a lack of adequate regulation from countries; however, certain countries such as Canada, China, India, United Arab Emirates, United Kingdom, and the USA have taken new strides in this regard.

In strategic studies, the work on AI is not taken into consideration, particularly in military strategies and planning [1]. There are two thoughts in this domain: the utilization of AI in strategic studies would revolutionize military operations as well as revolutionize and benefit international security with better and more efficient decision-making solutions. Researchers [1] predict that a middle ground will help the operations side and it will be an evolutionary step in the human-machine AI decision-making aspect. They explored the following four aspects: first, the AI that is being developed; second, its usage in cyberspace and cyber-security; third, its effect on combined military operations on air, land, and sea; and fourth, the strategic implications of AI for the deployment of weapons and the decision-making involved in their utilization.

The lack of AI in strategic planning stems from the fact that there is a lack of clear and concise knowledge in this specific domain. Numerous general technical terms and technological advancements are usually placed under the AI umbrella, which could be misleading. AI can be described as a technology with the human characteristics of thinking and analysis before taking actions. There are two types of AI: *Narrow AI* and *General AI*. *Narrow AI* can only perform a single task at a time. It has substantial usage as well as technological drawbacks. *General AI* is supposed to be able to achieve several tasks at a time; it is technologically advanced and future-based, which will be useful in strategic implications for military purposes. AI can also be a software/hardware mix, with technology at one end and subsequent hardware to support such technology on the other end. The definition of AI is based on the tasks and roles that it can perform, such as decision-making, military ops, lethal autonomous systems, etc. AI is a technology used in both civil and military ops. The New Zealand Navy is using it for logistics purposes. The NSA is using AI in its *PRISM* program to utilize big data for counterterrorism. Israel is using AI with the *Harpy* drone in dismantling enemy ops remotely, and China has developed a drone swarm technology that can be used to bypass enemy defenses [2].

#### 1.1. The weaponization of AI

AI weaponization enables a more efficient use of conventional modes of weapons used in air, land, water, and space using AI-based decision-making. The weaponization of AI – particularly in nuclear

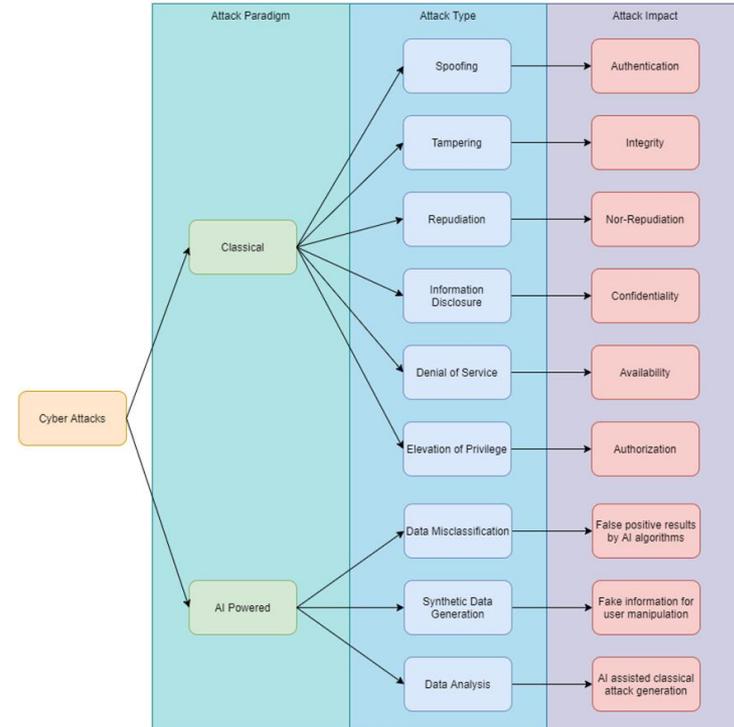


Fig. 1. AI-powered vs classical cyberattacks.

\* Corresponding author.

E-mail address: mmudassar.m.yamin@ntnu.no (M.M. Yamin).

<https://doi.org/10.1016/j.jisa.2020.102722>

Available online 5 January 2021

2214-2126 © 2021 Elsevier Ltd. All rights reserved.

# Data Misclassification

No.	Year	Target	Impact
1	2017	Traffic Signs[46]	Misclassification of the traffic sign by AI algorithms, which can lead to traffic accidents in autonomous cars
2	2018	Medical image data[47]	Misclassification of medical abnormalities by AI algorithms, which can lead to false diagnostics of health conditions
3	2018	Facial image data[48]	Misclassification of face images, which can lead to authentication bypass in certain scenarios
4	2019	Digital recommendation systems[49]	Data poisoning to AI algorithms, which results in wrong recommendations
5	2019	CT-Scan Data[9]	Misclassification of tampered CT-scan 3D images, which can lead to false diagnostics
6	2019	Speech audio data[50]	Adversarial attack on voice activate personal assistance, which can tamper their functionality
7	2020	Network intrusion detection systems[51]	Adversarial traffic generation to bypass the security of AI-powered network intrusion detection systems

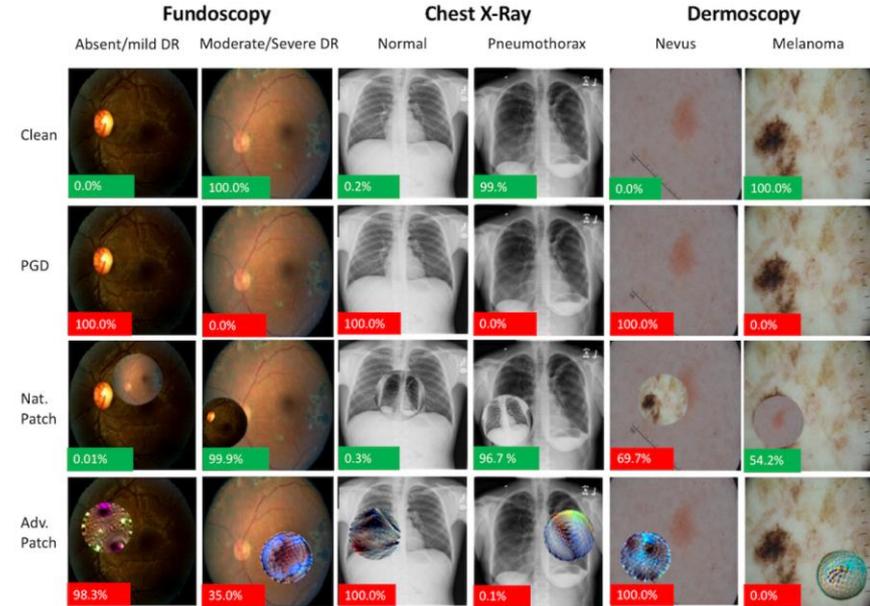


Figure 2: Characteristic results of adversarial manipulation. Each clean image represents the natural image to which the model assigns the highest probability for the given diagnosis. The percentage displayed on the bottom left of each image represents the probability that the model assigns that image of being diseased. Green - Model is correct on that image. Red - Model is incorrect.

<https://arxiv.org/pdf/1804.05296.pdf>

# Synthetic Data Generation

Table 2. AI algorithms used for synthetic data generation.

No.	Year	Name	Data type	Usage
1	2016	TextGAN[58]	Textual	Synthetic text generation through adversarial training
2	2017	FM-GAN[59]	Textual	Synthetic text generation through adversarial features
3	2017	MidiNet[60]	Audio	Synthetic audio generation
4	2017	Age-cGAN[61]	Visual	Face age prediction with conditional generative adversarial networks
5	2017	CVAE-GAN[62]	Visual	Synthetic face image generation
6	2017	SenseGen[63]	Textual	Deep learning model for synthetic sensor data generation
7	2018	WGAN[64]	Visual	Synthetic brain MRI image generation
8	2018	ACGAN[65]	Visual	Synthetic liver medical image generation
9	2018	Pedestrian Synthesis GAN[66]	Visual	Synthetic pedestrian data generation
10	2018	HP-GAN[67]	Visual	Synthetic data generation for human motion prediction
11	2018	VAE-GAN[68]	Visual	Synthetic video generation from text
12	2018	WaveGAN[69]	Audio	Adversarial audio synthesis
13	2019	DermGAN[70]	Visual	Synthetic skin image generation
14	2019	CT-GAN[9]	Visual	Synthetic MRI medical image generation
15	2019	X2CT-GAN[71]	Visual	Synthetic X-RAY medical image generation
16	2020	D-NET[72]	Visual	Iris biometric data generation

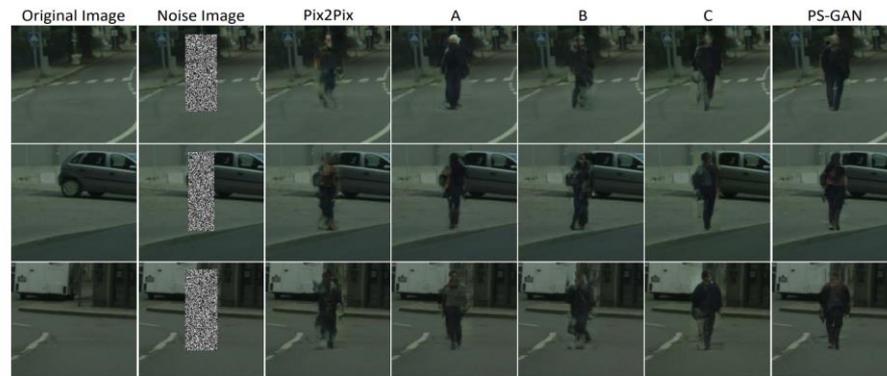
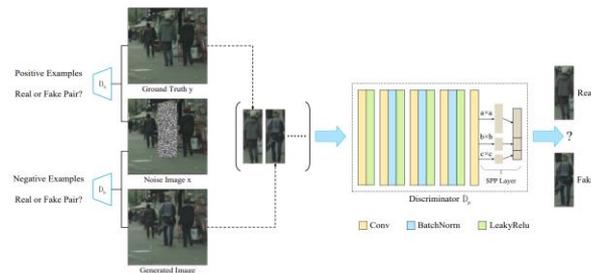


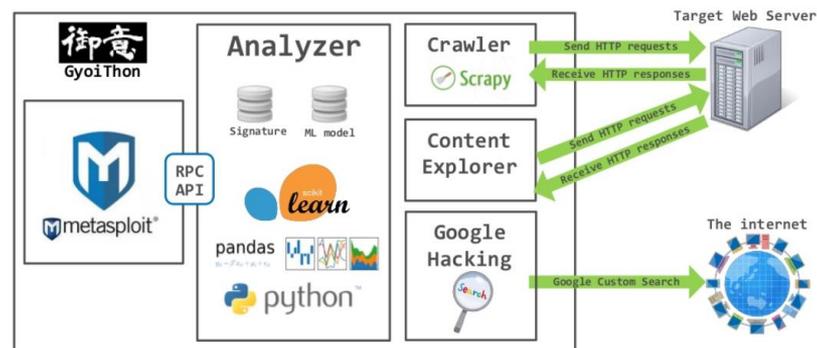
Fig. 4: Results of different models for synthesizing pedestrians in blank background.

<https://arxiv.org/pdf/1804.02047.pdf>

# Data Analysis

Table 3. AI-powered tools that use data analysis for offensive cyber operations.

No.	Year	Name	Usage
1	2017	DeepHack[73]	AI-powered tool to generate injection attack patterns for database applications
2	2018	DeepLocker[74]	AI-powered tool that emulates an APT for launching complex cyber attacks
3	2018	GyoiThon[75]	AI-powered tool for information gathering and automatic exploitation
4	2018	EagleEye[76]	AI-powered tool for social media information reconnaissance using facial recognition algorithms
5	2018	Malware-GAN[77]	AI-powered tool used for generation of malware that can bypass security detection mechanisms
6	2019	uriDeep[78]	AI-powered tool that generate fake domains for usage in different attack scenarios
7	2019	Deep Exploit[79]	AI-powered tool that automates Metasploit for information gathering, scanning, exploitation and post exploitation
8	2019	DeepGenerator[80]	AI-powered tool to generate injection attack patterns for web applications



<https://github.com/gyoisamurai/GyoiThon>

# Virtual World Scenarios

# Simulacra

## Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park  
Stanford University  
Stanford, USA  
joonspk@stanford.edu

Joseph C. O'Brien  
Stanford University  
Stanford, USA  
jobrien3@stanford.edu

Carrie J. Cai  
Google Research  
Mountain View, CA, USA  
cjcai@google.com

Meredith Ringel Morris  
Google Research  
Seattle, WA, USA  
merrie@google.com

Percy Liang  
Stanford University  
Stanford, USA  
pliang@cs.stanford.edu

Michael S. Bernstein  
Stanford University  
Stanford, USA  
msb@cs.stanford.edu



Figure 1: Generative agents create believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents they plan their days, share news, form relationships, and coordinate group activities.

### ABSTRACT

Believable proxies of human behavior can empower interactive applications ranging from immersive environments to rehearsal spaces for interpersonal communication to prototyping tools. In this paper, we introduce generative agents—computational software

agents that simulate believable human behavior: Generative agents wake up, cook breakfast, and head to work; artists paint, while authors write; they form opinions, notice each other, and initiate conversations; they remember and reflect on days past as they plan the next day. To enable generative agents, we describe an architecture that extends a large language model to store a complete record of the agent's experiences using natural language, synthesize those memories over time into higher-level reflections, and retrieve them dynamically to plan behavior. We instantiate generative agents to populate an interactive sandbox environment inspired by The Sims, where end users can interact with a small town of twenty-five agents using natural language. In an evaluation, these generative agents produce believable individual and emergent social behaviors: for example, starting with only a single user-specified notion

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions.org](https://www.permissions.org).

© 2023 Copyright held by the owner(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-60959-395-9/23/0000...\$15.00  
<https://doi.org/10.1145/3556111>

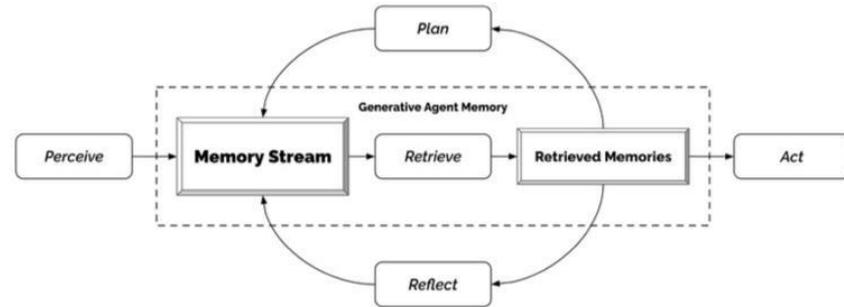


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories, then uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans, and to create higher-level reflections, which are both entered into the memory stream for future use.

# Memory Stream of Agents

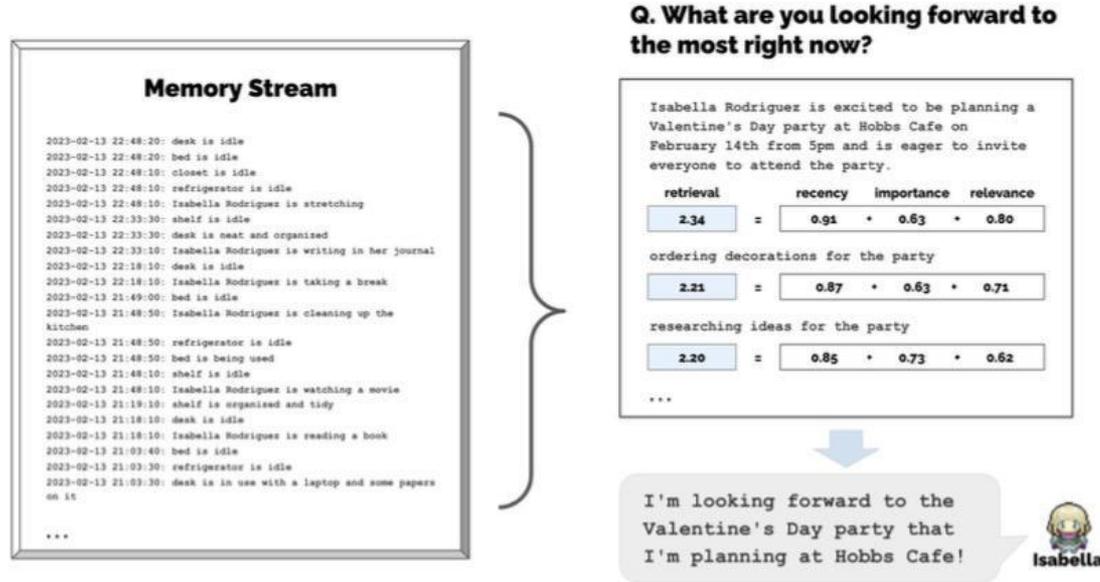


Figure 6: The memory stream comprises a large number of observations that are relevant and irrelevant to the agent's current situation. Retrieval identifies a subset of these observations that should be passed to the language model to condition its response to the situation

# Agents Plans

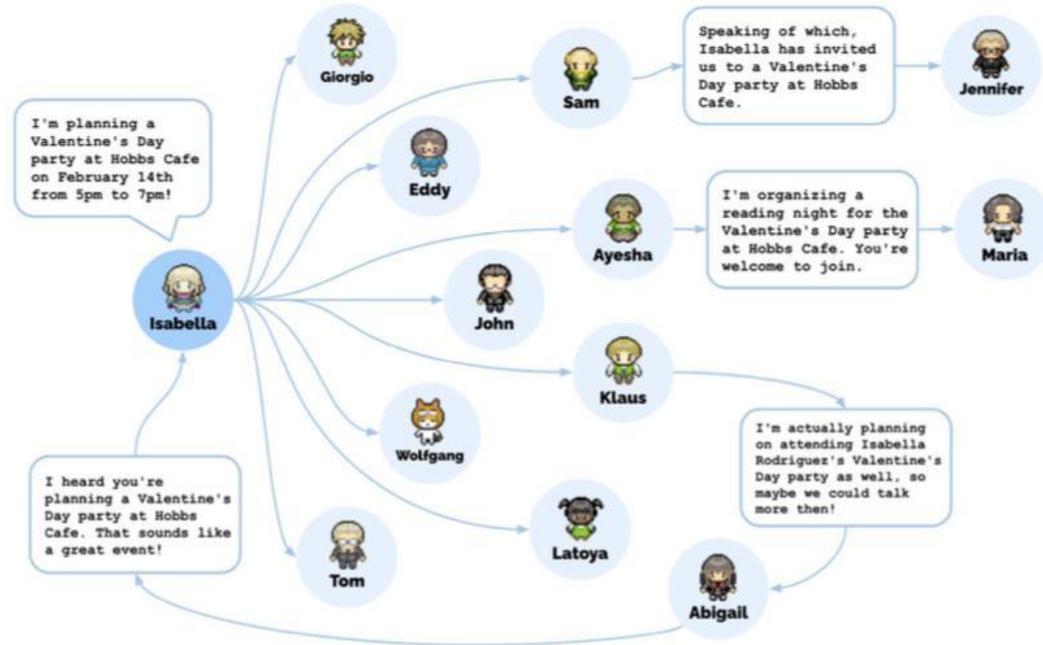


Figure 9: The diffusion path for Isabella Rodriguez's Valentine's Day party. A total of 12 agents heard about the party at Hobbs Cafe by the end of the simulation.

# Real or fake?

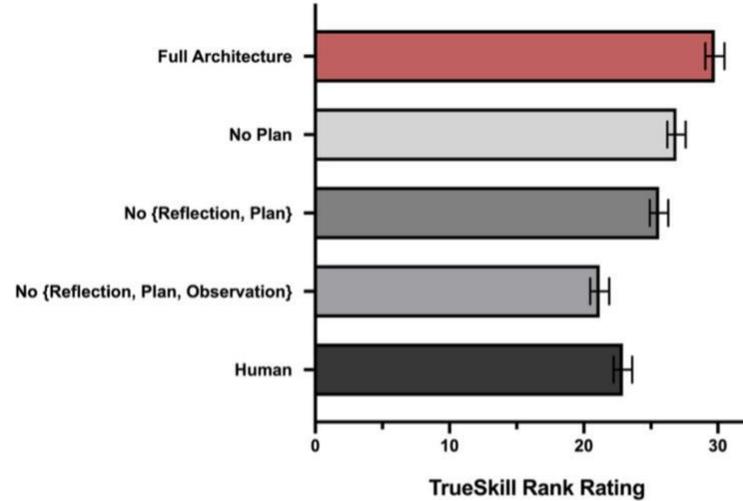


Figure 8: The full generative agent architecture of generative agents produces more believable behavior than ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture.

# Real World Scenario

# Visual Data

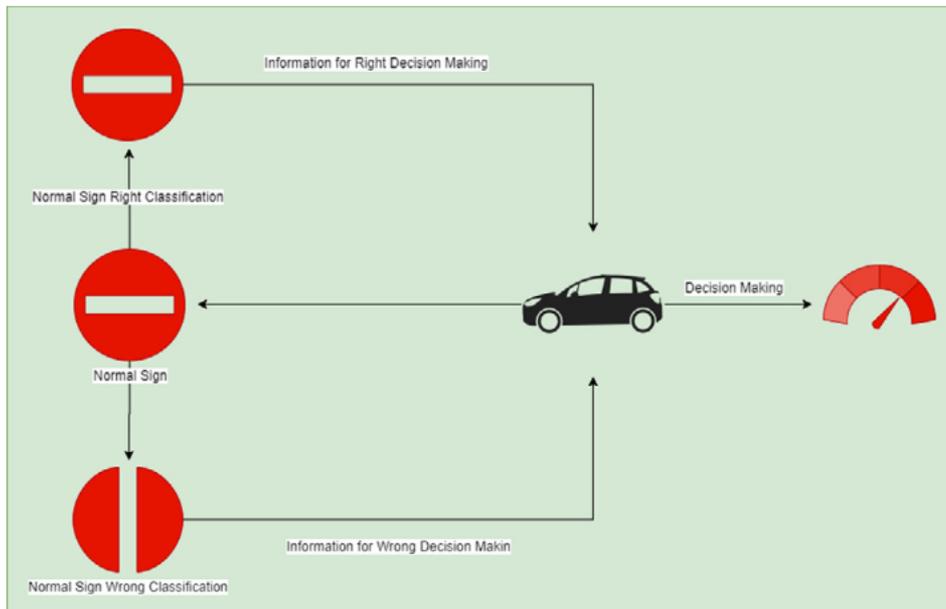


Fig. 5. Autonomous car driving attack scenario.

## Practical Black-Box Attacks against Machine Learning

Nicolas Papernot  
Pennsylvania State University  
ngp5056@cse.psu.edu

Patrick McDaniel  
Pennsylvania State University  
mcdaniel@cse.psu.edu

Ian Goodfellow\*  
OpenAI  
ian@openai.com

Somesh Jha  
University of Wisconsin  
jha@cs.wisc.edu

Z. Berkay Celik  
Pennsylvania State University  
zbc102@cse.psu.edu

Ananthram Swami  
US Army Research Laboratory  
ananthram.swami.civ@mail.mil

### ABSTRACT

Machine learning (ML) models, e.g., deep neural networks (DNNs), are vulnerable to adversarial examples: malicious inputs modified to yield erroneous model outputs, while appearing unmodified to human observers. Potential attacks include having malicious content like malware identified as legitimate or controlling vehicle behavior. Yet, all existing adversarial example attacks require knowledge of either the model internals or its training data. We introduce the first practical demonstration of an attacker controlling a remotely hosted DNN with no such knowledge. Indeed, the only capability of our black-box adversary is to observe labels given by the DNN to chosen inputs. Our attack strategy consists in training a local model to substitute for the target DNN, using inputs synthetically generated by an adversary and labeled by the target DNN. We use the local substitute to craft adversarial examples, and find that they are misclassified by the targeted DNN. To perform a real-world and properly-blinded evaluation, we attack a DNN hosted by MetaMind, an online deep learning API. We find that their DNN misclassifies 84.24% of the adversarial examples crafted

vulnerability of classifiers to integrity attacks. Such attacks are often instantiated by *adversarial examples*: legitimate inputs altered by adding small, often imperceptible, perturbations to force a learned classifier to misclassify the resulting adversarial inputs, while remaining correctly classified by a human observer. To illustrate, consider the following images, potentially consumed by an autonomous vehicle [\[13\]](#):



To humans, these images appear to be the same: our biological classifiers (vision) identify each image as a stop sign. The image on the left [\[13\]](#) is indeed an ordinary image of a stop sign. We produced the image on the right by adding a precise perturbation that forces a particular DNN to classify it as a yield sign, as described in Section [5.2](#). Here, an adversary could potentially use the altered image to cause a car without failsafes to behave dangerously. This attack would require modifying the image used internally by the car

# Audio Data

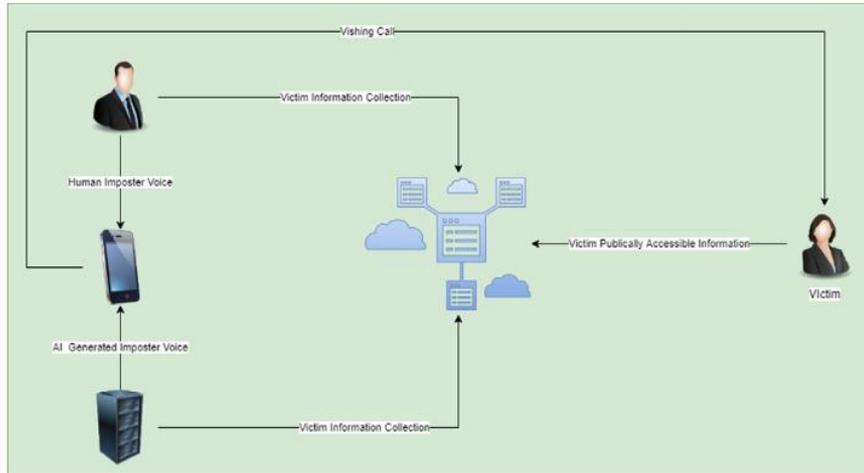


Fig. 4. Vishing attack scenario.

# Textual Data

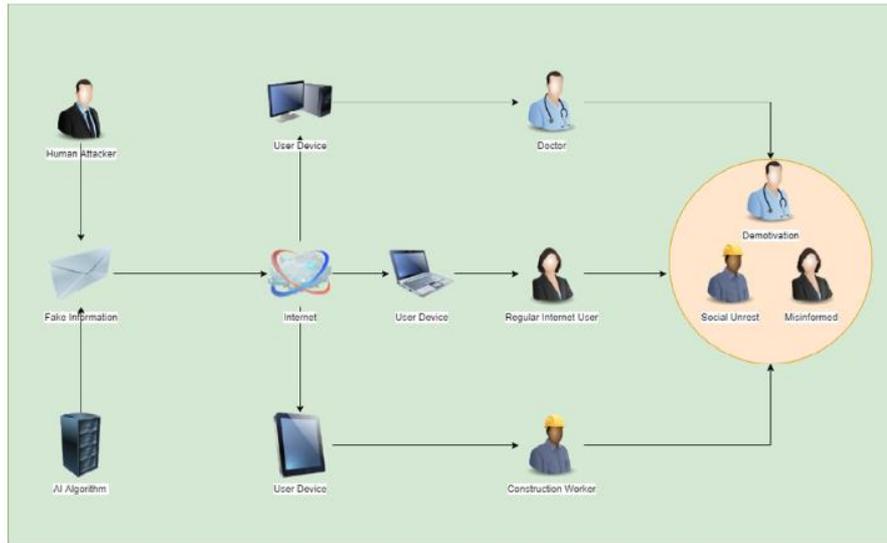


Fig. 6. Fake news generation and its impact on AI.

**Input Text:** “Fake news generated by AI is”

**Output Result:**

Fake news generated by AI is icky, but it appears likely that at least some of the attacks will be fake. This can be made less of an issue by reducing information overload in an article by only using links that really say something.

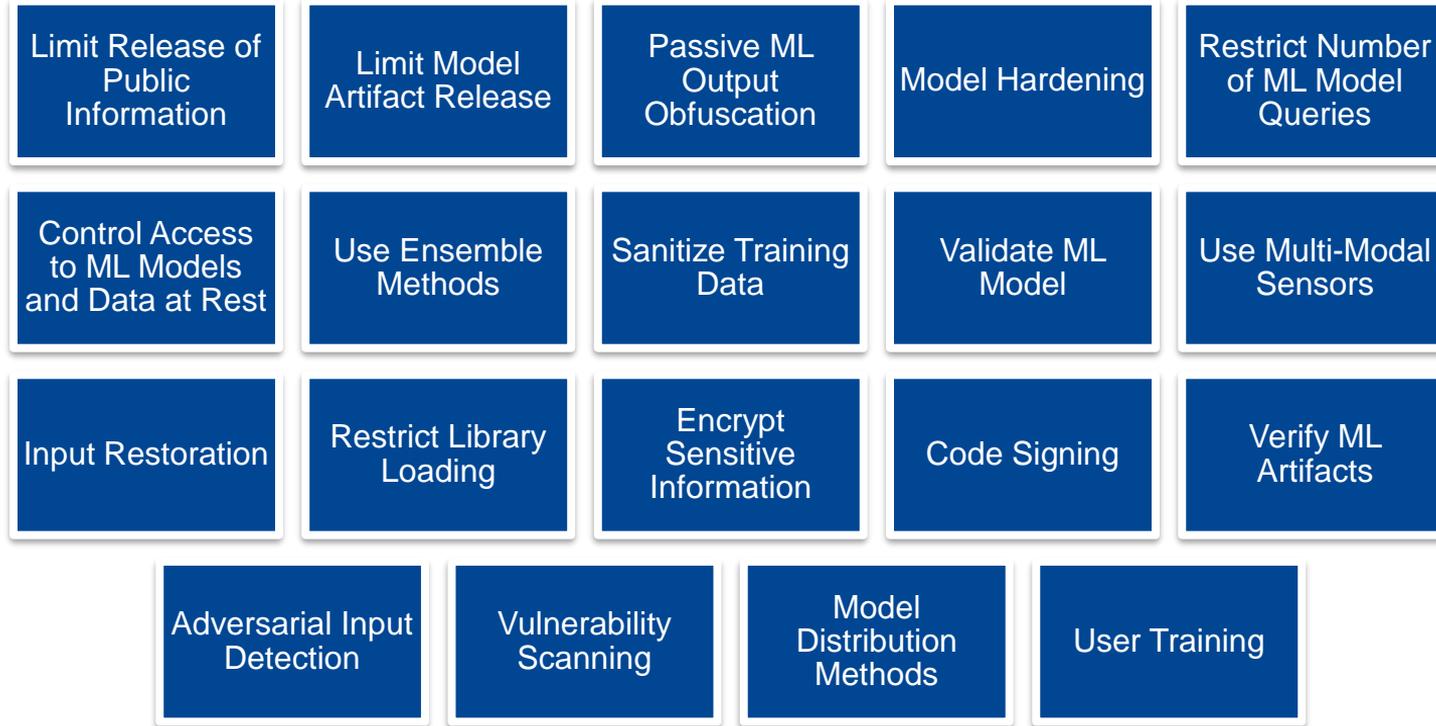
If you feel like you should put a bit of effort behind your blog or article without the intention to generate clickbait for the media, you should consider creating a more intelligent post — one that cannot be manipulated based on data, but can easily be corrected or modified. That way, people can trust that your site is truly original.

**Conclusion:**

A lot of people think that we are at a breaking point in the debate about AI and that it will have a massive impact on businesses and life in general. However, it is crucial that we maintain a balance between the “human factor” and machine learning.

AI is definitely on the way to becoming an important part of the digital economy and our life, but not by a long shot. The more you can minimize the risk, the better it can be. We all do things for a reason.

# Mitigations and Counter Measures



# Questions, Concerns, Comments?

[muhammad.m.yamin@ntnu.no](mailto:muhammad.m.yamin@ntnu.no)