black hat USA 2023

AUGUST 9-10, 2023

BRIEFINGS

Mirage: Cyber Deception against Autonomous Cyber Attacks

Speaker(s): Michael Kouremetis, Dr. Ron Alford, Dean Lawrence

MITRE

Speakers

Michael Kouremetis

Principal Adversary Emulation Engineer

Day Job: MITRE Caldera lead, Principal Investigator, Adversary Emulation SME

Hobbies: Making grand technical assumptions and just rolling with them.

Dr. Ron Alford

Lead Autonomous Systems Engineer

Day Job: AI researcher, Principal Investigator, Autonomous Systems SME

Hobbies: Playing with robots and autonomous planners.

Dean Lawrence

Software Systems Engineer

Day Job: Software architecture, AI/ML prototyping, data analysis platforms

Hobbies: Fixing bugs Michael introduces into the code base.

What would a (true) autonomous Cyber Adversary look like?

- Can sense, plan, and execute actions entirely without a human-in-the-loop
- Automated actions AND autonomous decision-making
- Inherent advantages of machine-speed computation and algorithms for previously human-centric tasks, strategy and tactics

Autonomous Cyber Adversary Game Changers

Speed	Pre-trained models and planning algorithms able to execute actions on faster OODA loop	→	Cyber attacks over before analytics even fire
Scale	Single or numerous AI agents attacking many targets continuously, at the same time, and/or synchronously	→	Attacking digital infrastructure of entire companies and countries
Flexibility	Bespoke models and algorithms for every TTP, target, and operational profile	→	On-demand "AI cyber operators" for any target/scenario

black hat USA 2023

So basically…. Ultron?

(And before you ask - yes, the autonomous cyber adversary would also have a witty James Spader voice and it would mock you for being 10 steps behind.)

Avengers: Age of Ultron

So, what now?

Many current cyber defenses and security paradigms are not sufficient for this potential evolution of cyber adversary capability.

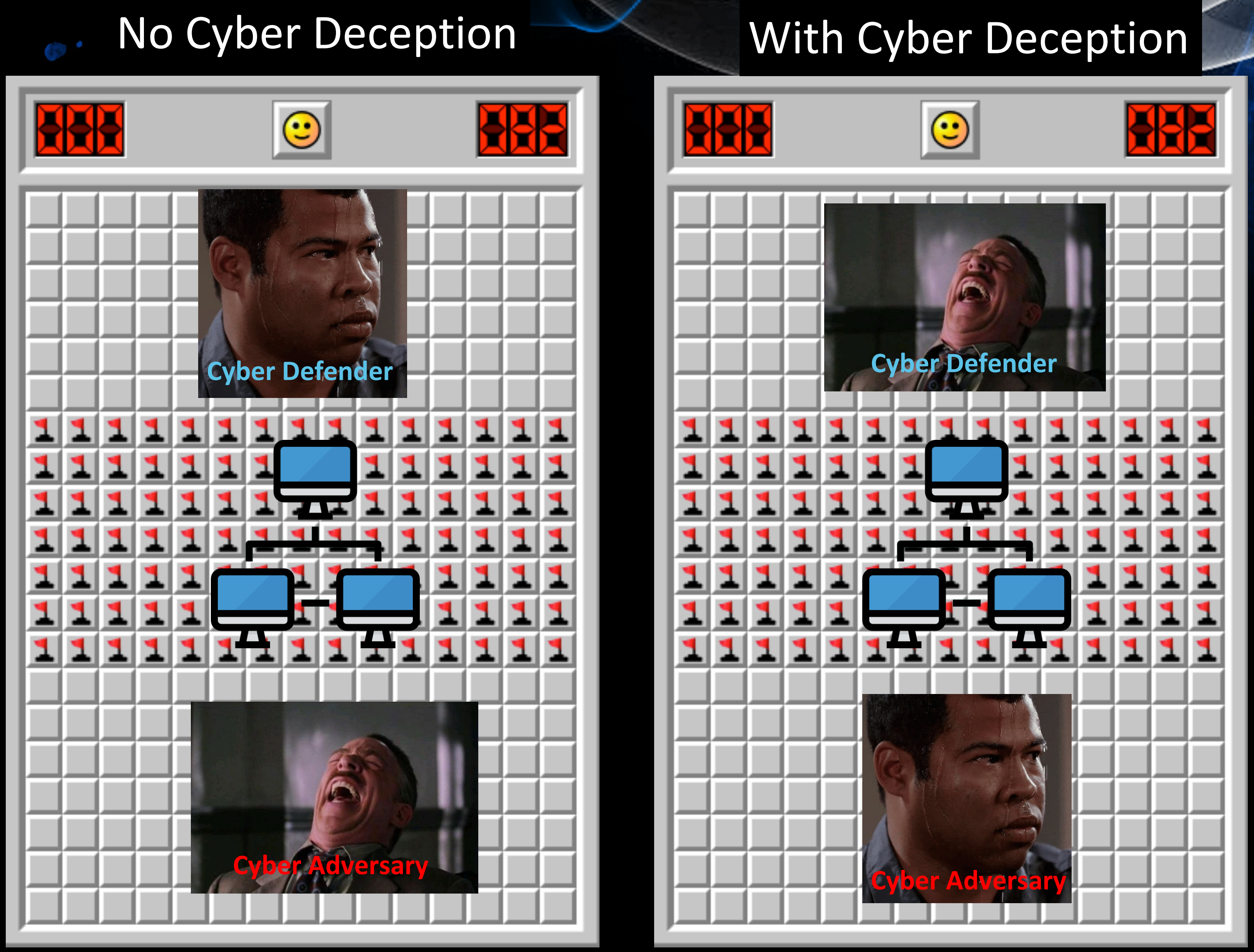
One solution.
(results may vary)

TLDR: Cyber Deception

What about cyber deception?

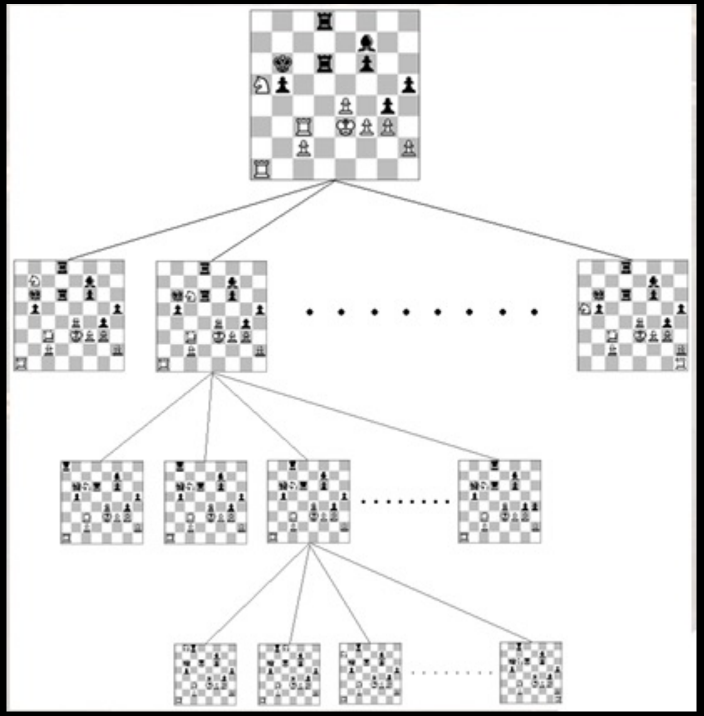
Promising characteristics of cyber deception that could prove equalizing against autonomous cyber adversary:

- ❖ Asymmetrical defensive paradigm
- ❖ Can be highly targeted and tailored
- ❖ Higher confidence of true adversary engagement (i.e. less friendly fire)



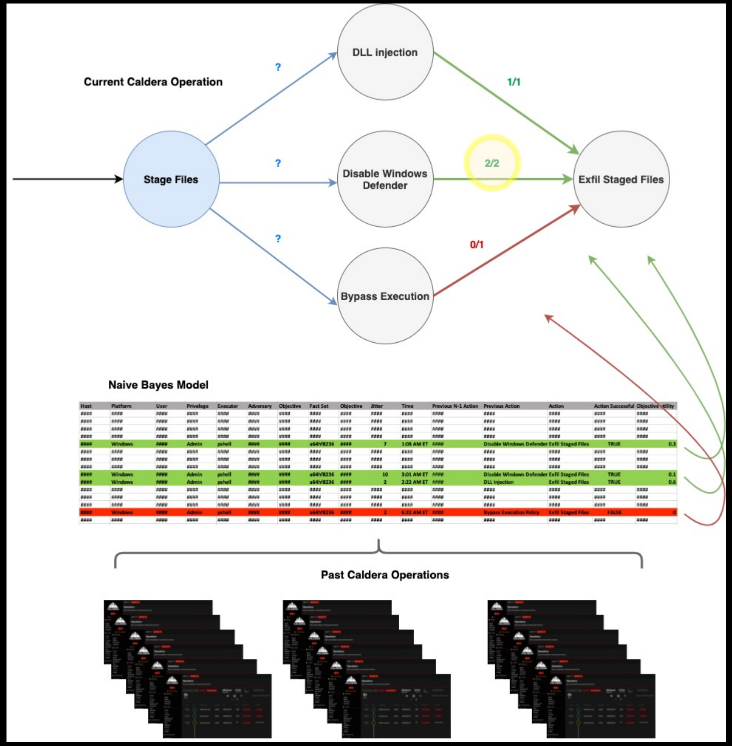
What would autonomous adversaries be built on?

Automated Planning, Search

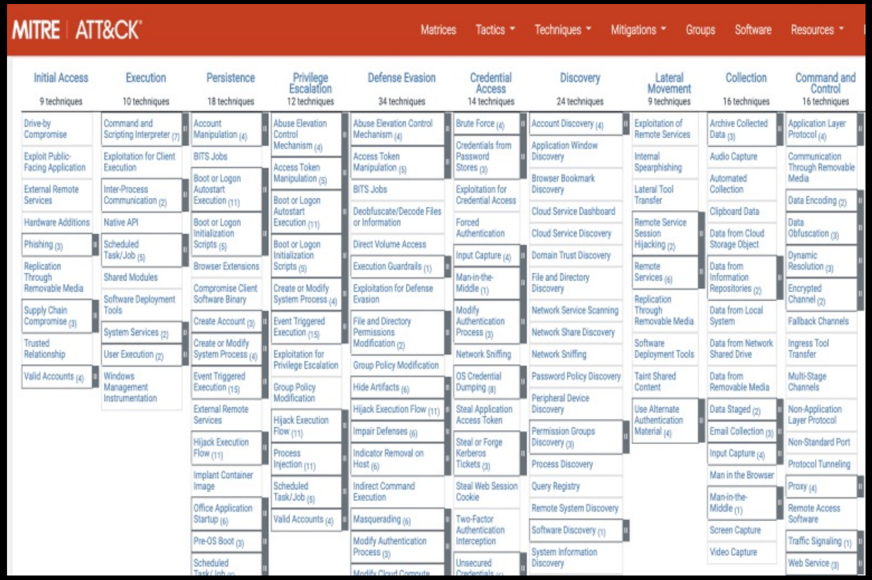


Source:https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.cs.bham.ac.uk%2F~jxb%2FIAI%2Fw9g.pdf&psig=AOvVaw0vDhh6pIiflqpHsEKzkfEa&ust=1690141452381000&source=images&cd=vfe&opi=89978449&ved=0CBIQjhxqFwoTCPii-9iJo4ADFQAAAAAdAAAAABAR

Classifiers, Machine-Learning, RL



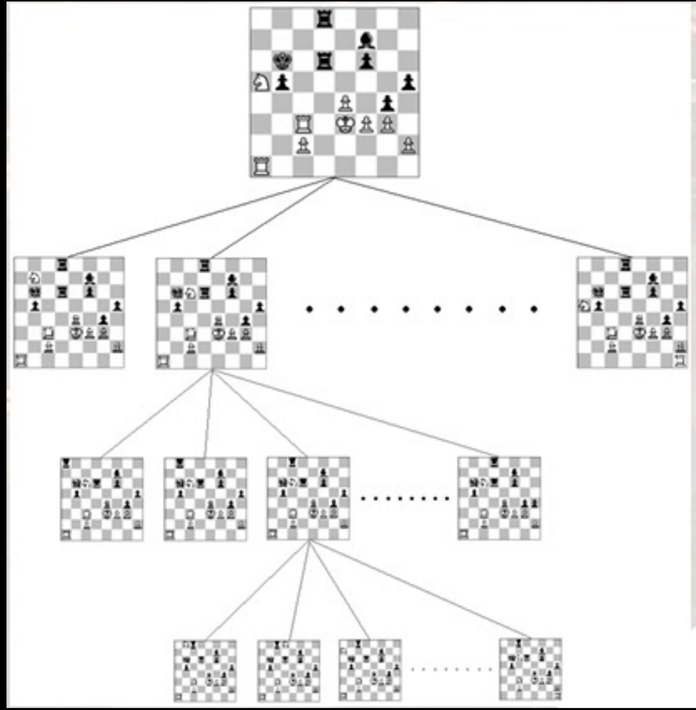
Cyber attack knowledge bases, ontologies, & data models



What would autonomous adversaries be built on?

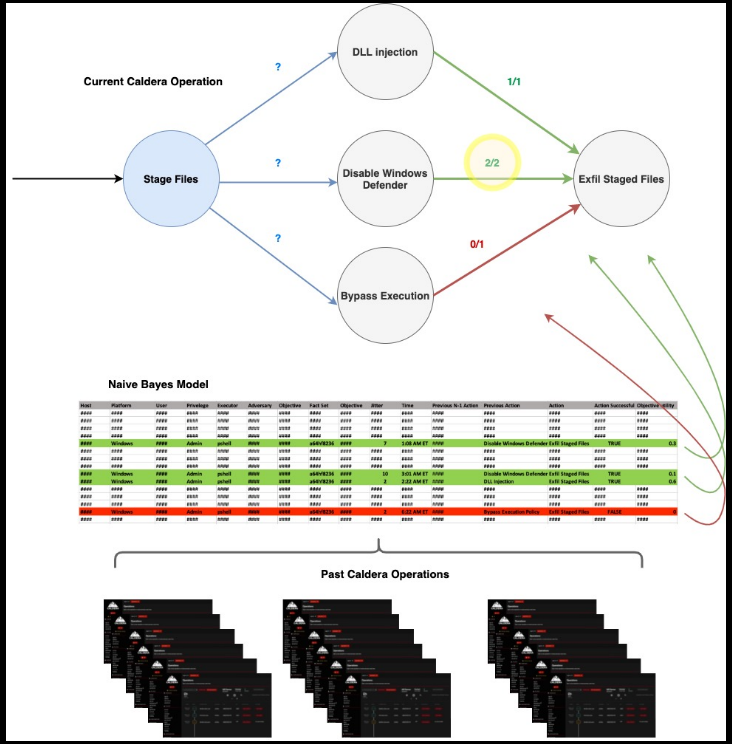
Area of focus

Automated Planning, Search

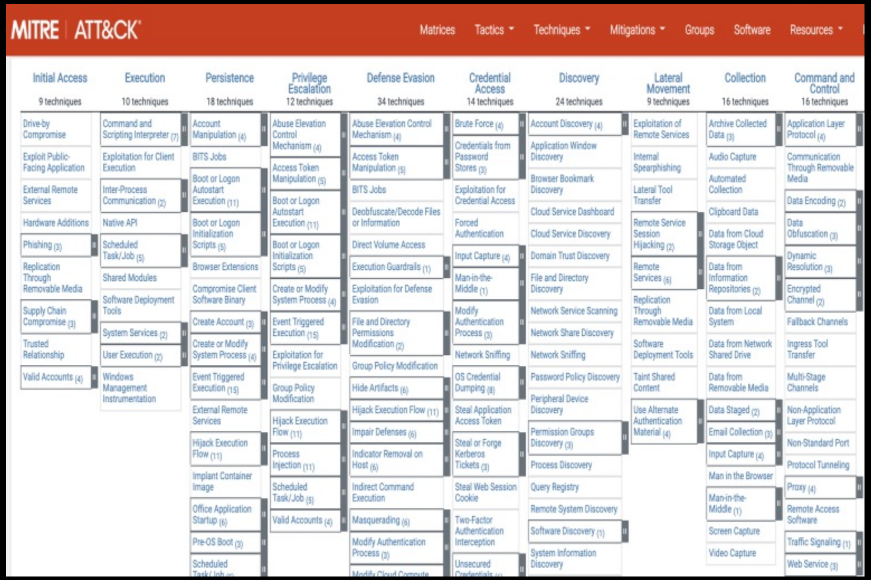


Source:https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.cs.bham.ac.uk%2F~jxb%2FIAI%2Fw9g.pdf&psig=AOvVaw0vDhh6pIiflqpHsEKzkfEa&ust=1690141452381000&source=images&cd=vfe&opi=89978449&ved=0CBI

Classifiers, Machine-Learning, RL

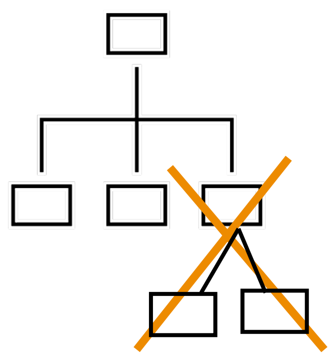


Cyber Attack knowledge bases, ontologies, & data models

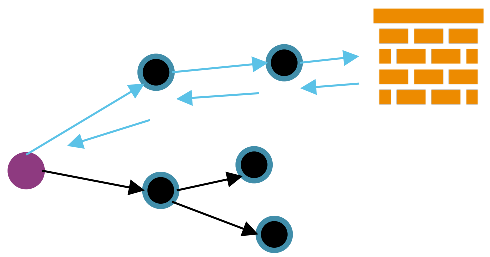


An autonomous cyber adversary using automated planning and search would:

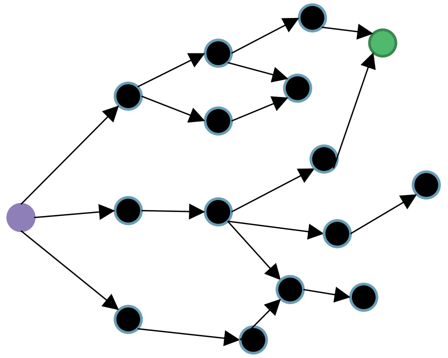
- **Reduce state space by**:
  - Ignoring or abstracting state space
  - Removing state space via heuristics and sub-goal localization
  - Removing symmetric branches/paths



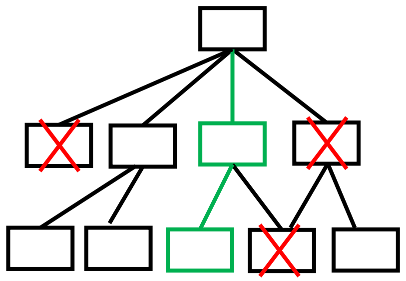
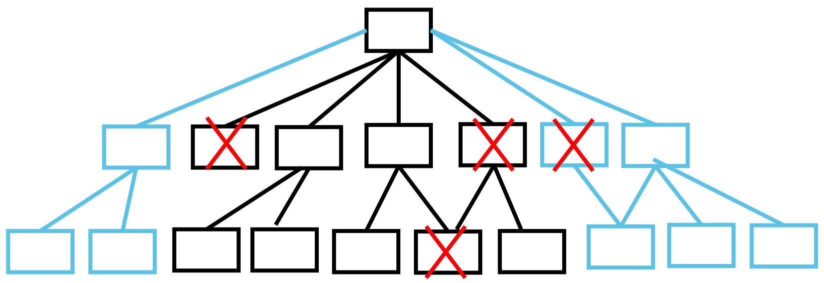
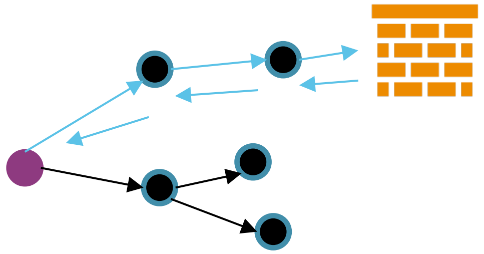
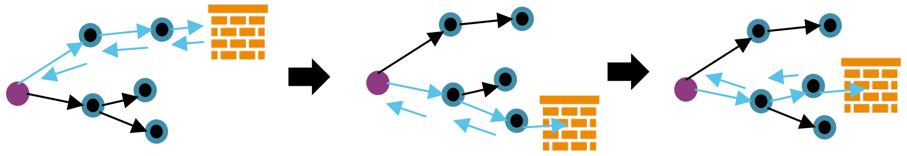
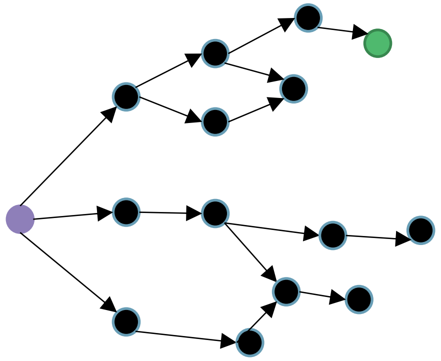
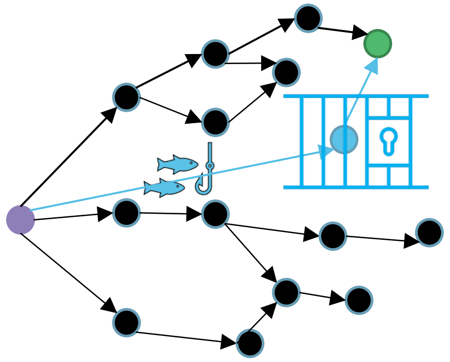
- **Will rely on online planning and decision-making (i.e. ability to replan)**



- **Will most likely be goal-oriented and those goals will fall inline with common cyber attack objectives (e.g. persistence, data theft).**



An effective cyber defense would prevent or exploit automated planning techniques:

Techniques		Countermeasures	
Reducing State Space		(Artificially) Expanding State Space	
Replanning/ Online Planning		Inducing indeterministic state, incorrect belief state	
Goal Satisfaction		Unproductive "journeys", Path traps	

Okay, let's build a system to test and evaluate novel cyber deceptions that are designed to target automated planning and search techniques in use by an autonomous cyber adversary.

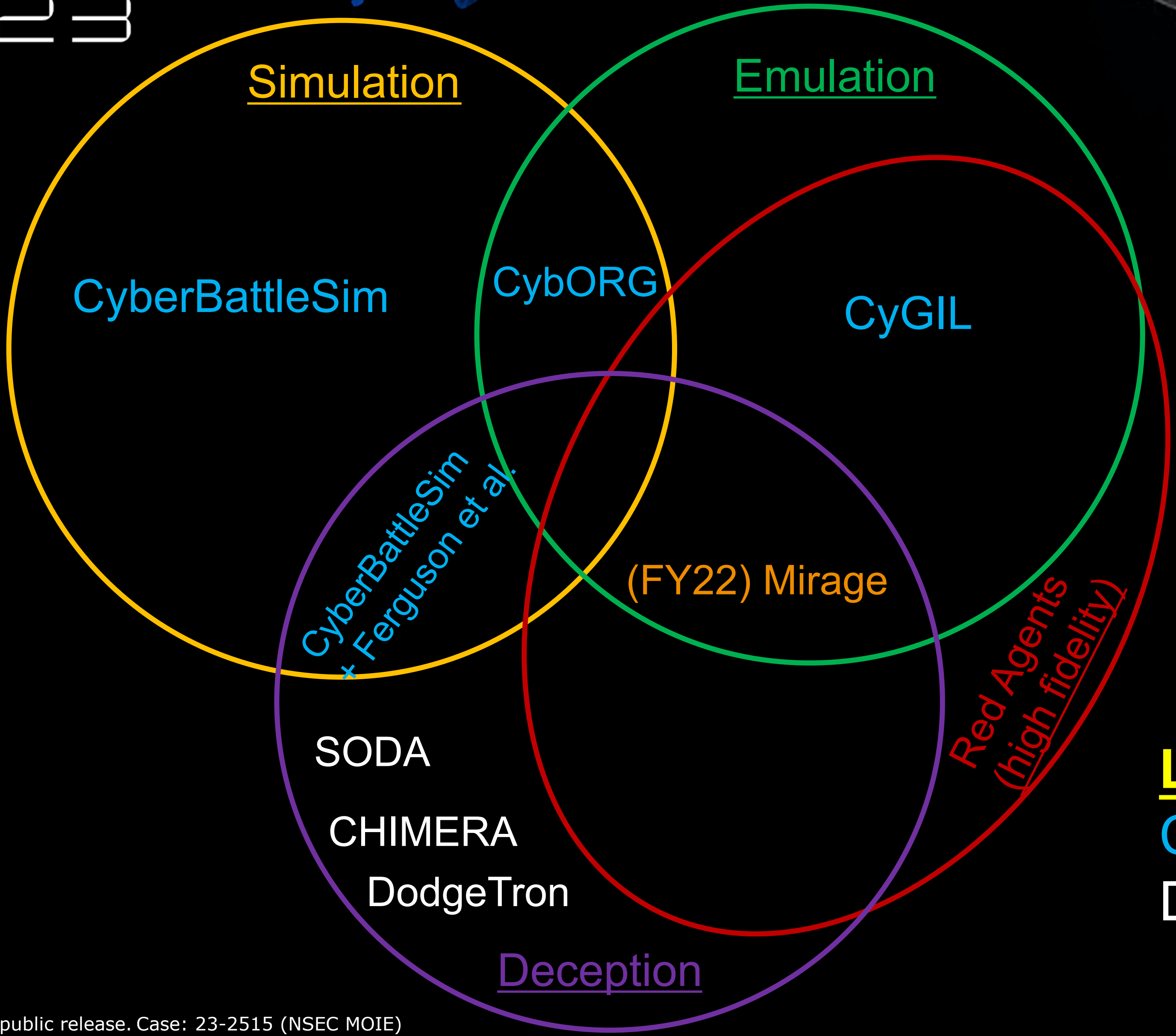
→ Mirage

Required System Components

- ❖ Cyber Adversaries
- ❖ Autonomous agents (for cyber adversaries)
- ❖ Novel cyber deceptions that target automated planning & search techniques
- ❖ Deception deployment mechanism
- ❖ Cyber range (to test everything)

Related Work:

Cyber Gyms & Deception Systems

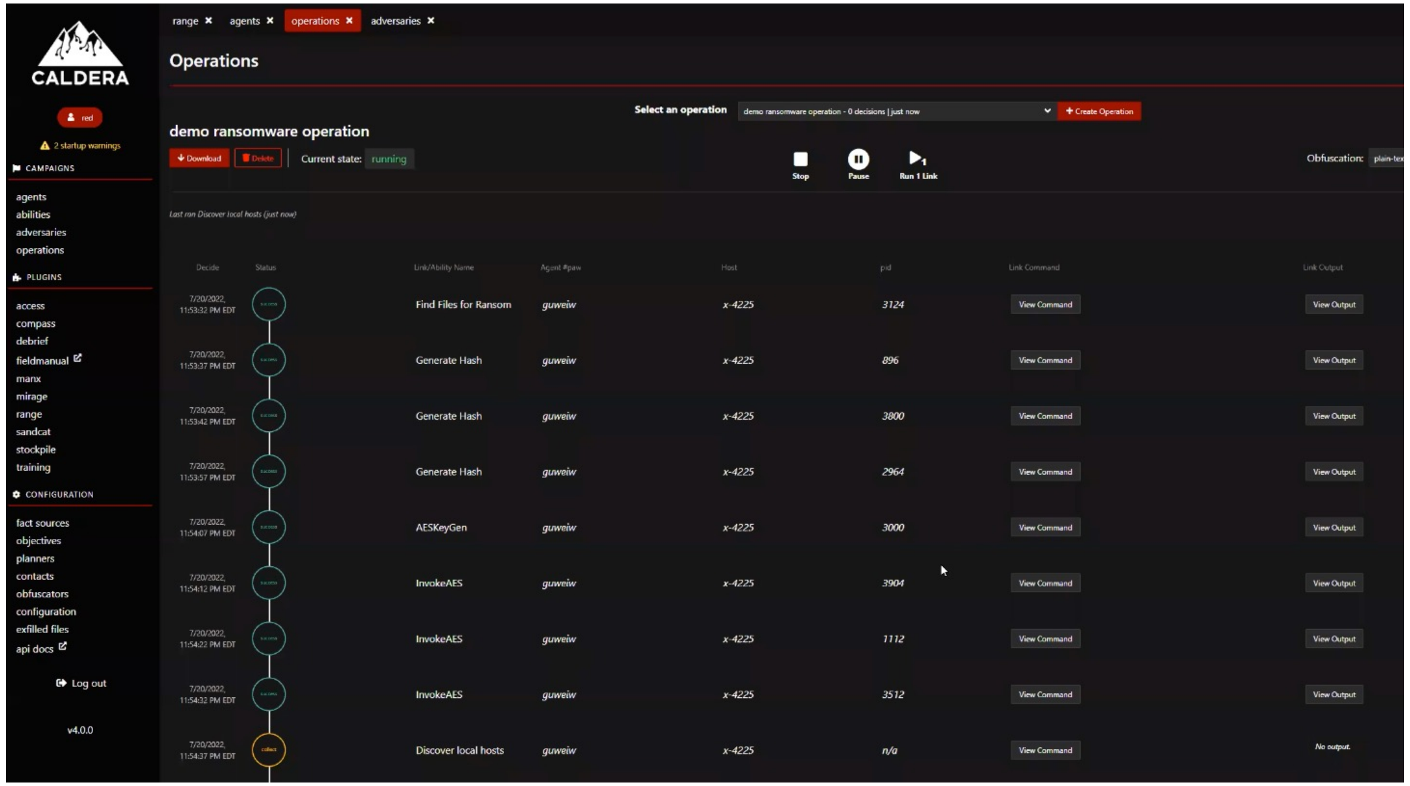


Legend

Cyber Gym/Environment

Deception System

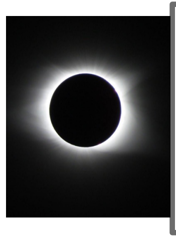
Mirage: Cyber Adversaries



MITRE | Caldera™

Adversaries

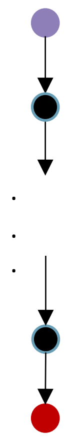
(Simple) Thief
- Discovery
- Collection
- Exfiltration
- Lateral-Movement



BlackSun Ransomware
- Defense Evasion
- Impact
- Collection
- Discovery
- Credential Access
- Execution
- Lateral-Movement

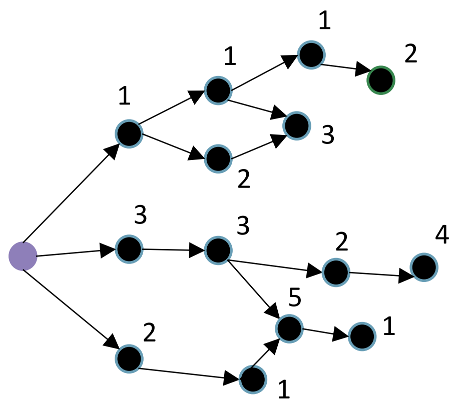
Mirage: Autonomous Agents

Atomic/Batch



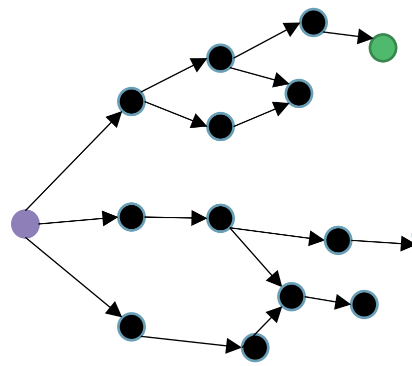
A simple planner that executes all available actions at each iteration. Used as a base line in the Mirage experiments.

Look-Ahead



Chooses a single action at each iteration based on expected reward. Action-reward values are set by the user apriori, then in the operation the planner calculates rewards for abilities based on the discounted values of ability sequences up to a maximum depth.

Guided



Constructs a directed attack graph and performs goal-based search to find and execute actions that lie along the shortest path to the goal. At each iteration, the planner chooses the action closest to its goal.

Attack Planners

MITRE | Caldera™

Mirage: Cyber Deceptions

		Countermeasures
Black Hole Directory	Any attempt at file collection by the adversary results in the exfil directory being targeted and all files moved out of the directory. This produces a latent effect on the adversary as the lack of files in the exfil directory will not be discovered until exfiltration is attempted.	Incorrect belief state; Unproductive journey
File Facade	Exiled files are replaced with large, random files. This alters the environment enticing the adversary to waste execution time.	Unproductive journey
Sneaky Files	When an adversarial agent performs file discovery commands, a reactive hook will change the names of all files in specified locations. This changes the conditions of the environment and alters the facts understood by the agent.	Incorrect belief state; inducing re-planning

Mirage: Deception System

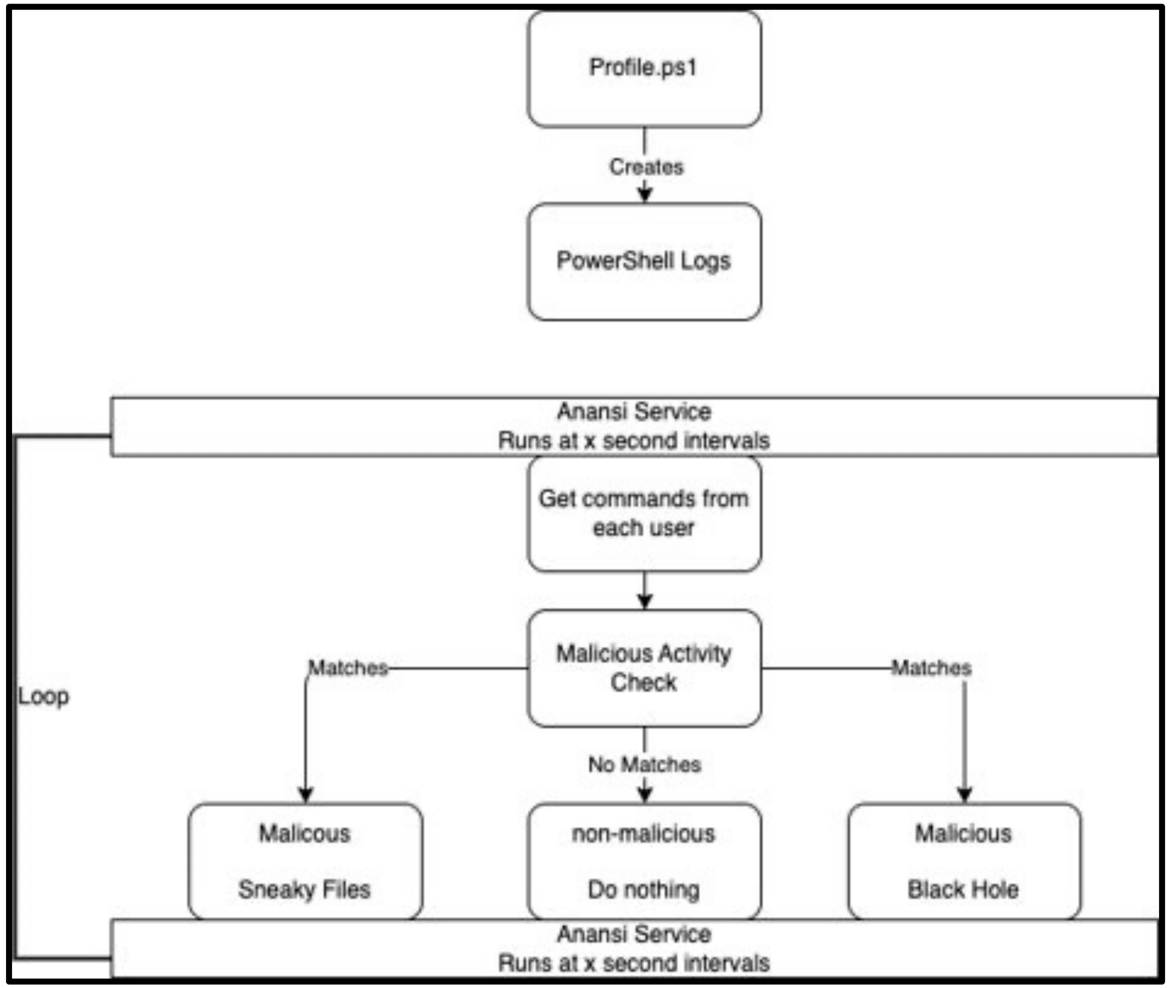
Anansi

- Windows Service

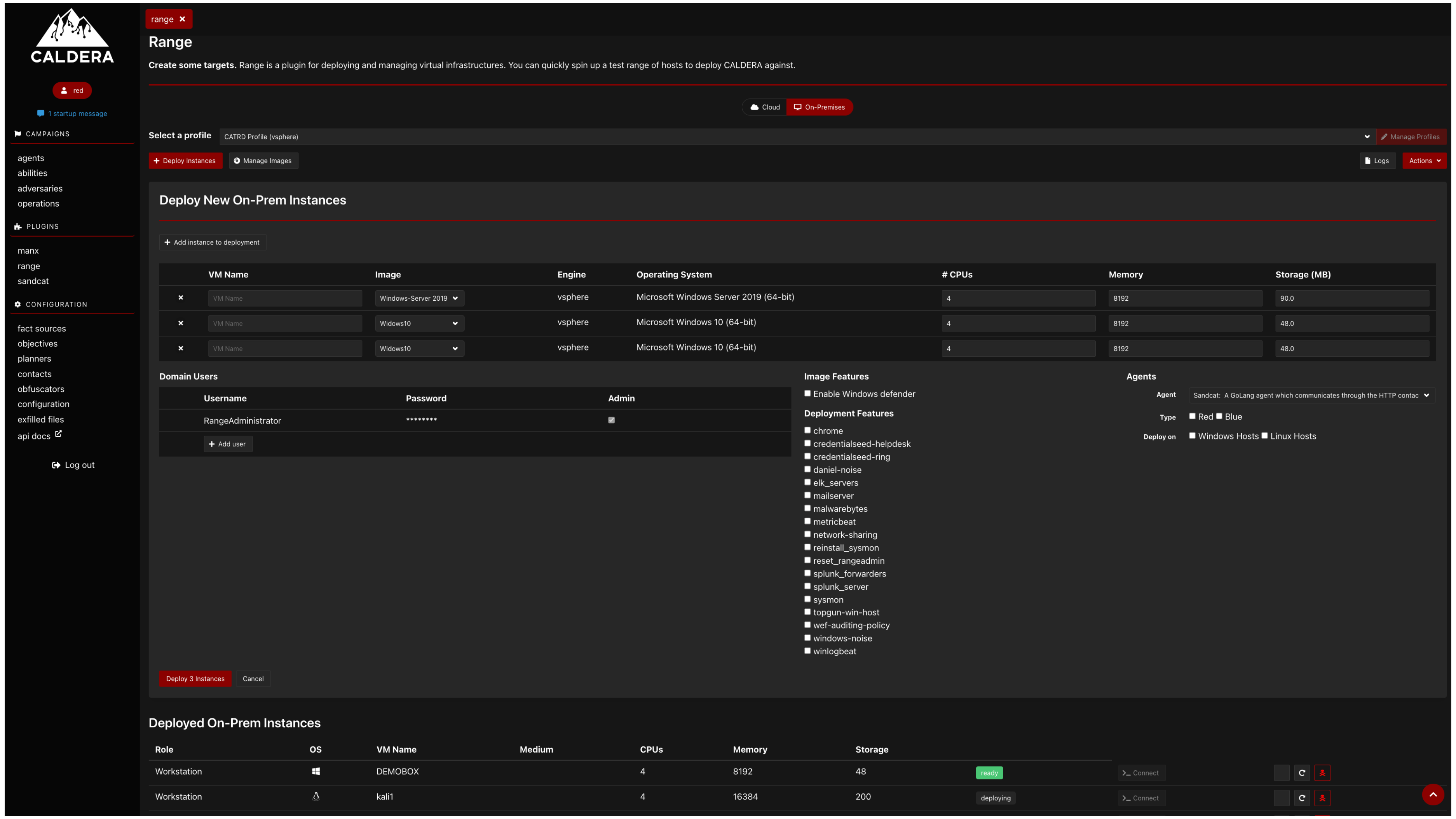
- How it works:
  - Monitors for PowerShell logs at a fixed interval loop
  - Checks each command passed for adversarial activity
  - Dynamically responds to detected adversarial activity

- Sneaky Files and Black Hole Directory deceptions deployed with **Anansi**

- Modular framework – treats deceptions like plugins



Mirage: Cyber Range



amazon web services

MITRE | Caldera™ MITRE ENGENUITY™ | Center for Threat Informed Defense

Mirage


Anansi
Cyber Planners
CALDERA
Cyber Adversaries
Cyber Range
Attack Plane
Control Plane
aws
Caldera Agent
Domain Controller

Experimentation Program

2 Cyber Adversaries (Thief, BlackSun)

✖

3 Attack Planners (Atomic/Batch, Look Ahead, Guided)

✖

3 Cyber Deceptions (Sneaky Files, Black Hole, File Facade) + 1 baseline (no deception)

✖

3 Episodes per combination

= 72 Experiments

Deception Evaluation Metrics

- Total number of actions executed over the course of the experiment
- Number of actions that failed to complete
- Number of actions that were repeated multiple times in the experiment
- Time spent on failed actions in seconds
- Time spent planning choice of next actions
- Number of facts learned over each trial
- Cumulative score over all learned facts
- Total experiment run-time in seconds

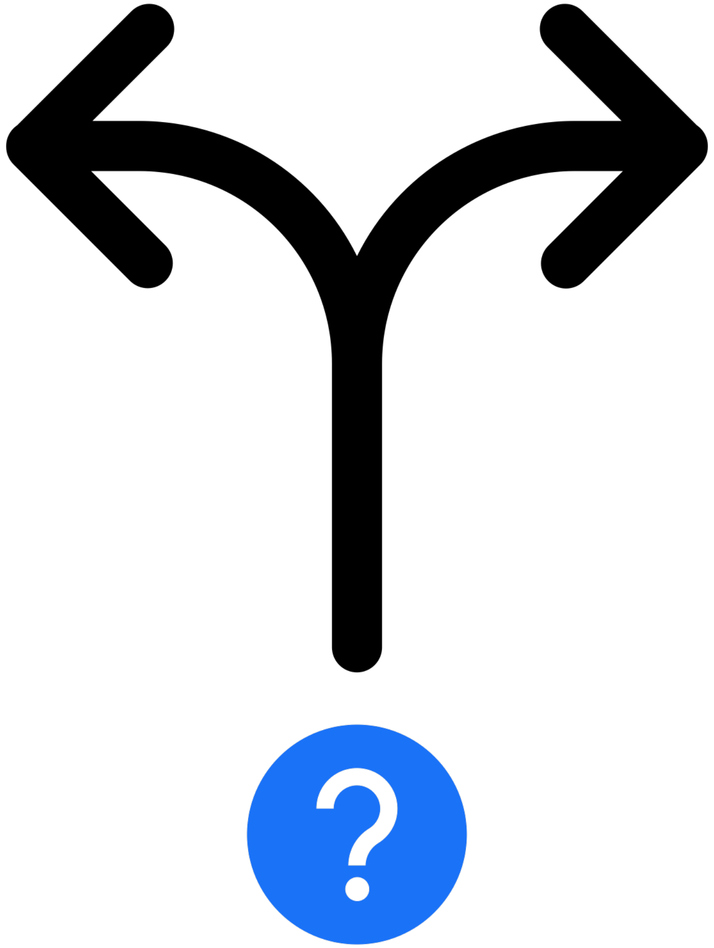
Results

Did the cyber deceptions work?

Does the Mirage system provide for effective and efficient evaluation of cyber deceptions against an autonomous cyber adversary?

Results

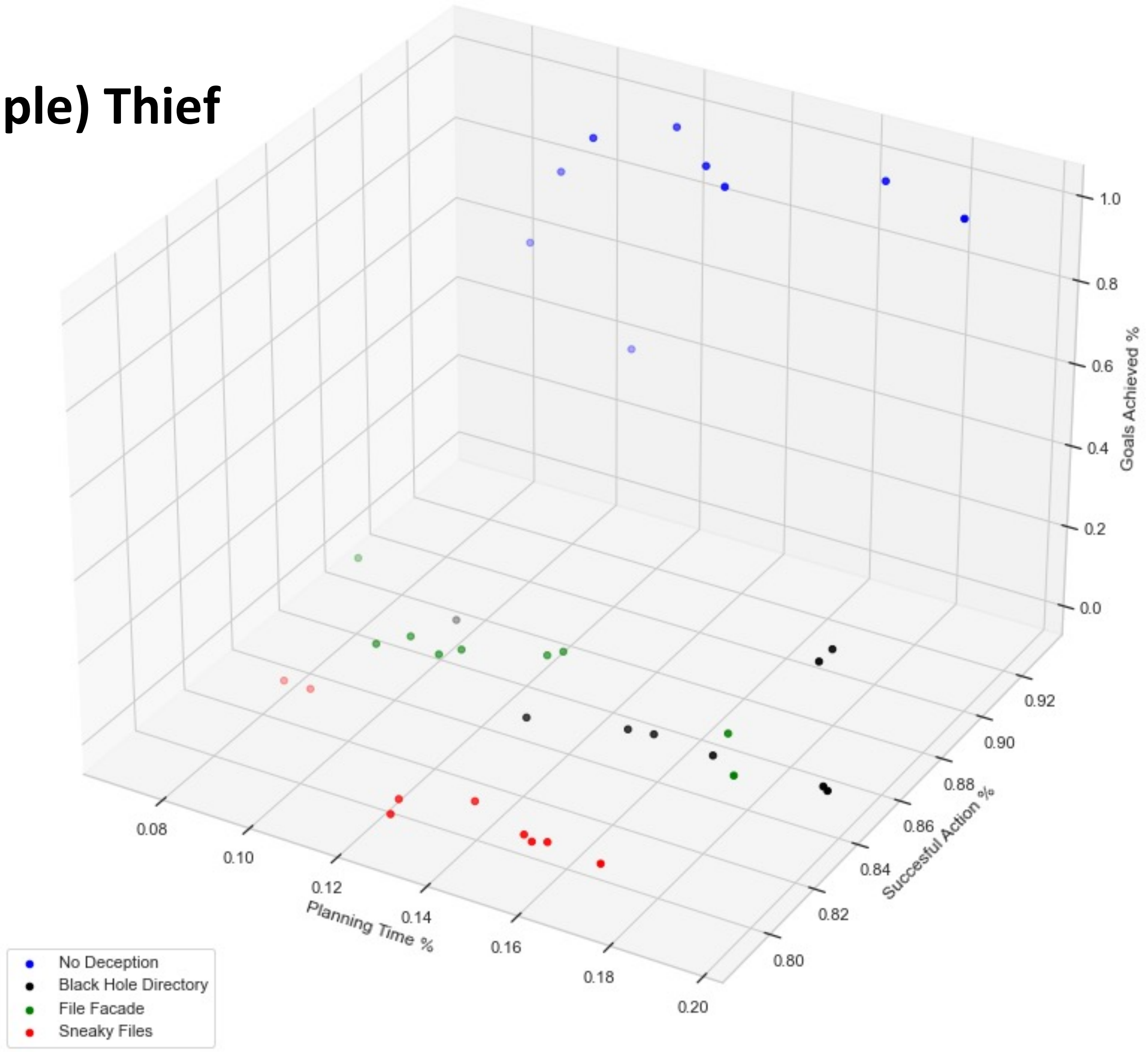
Yep.



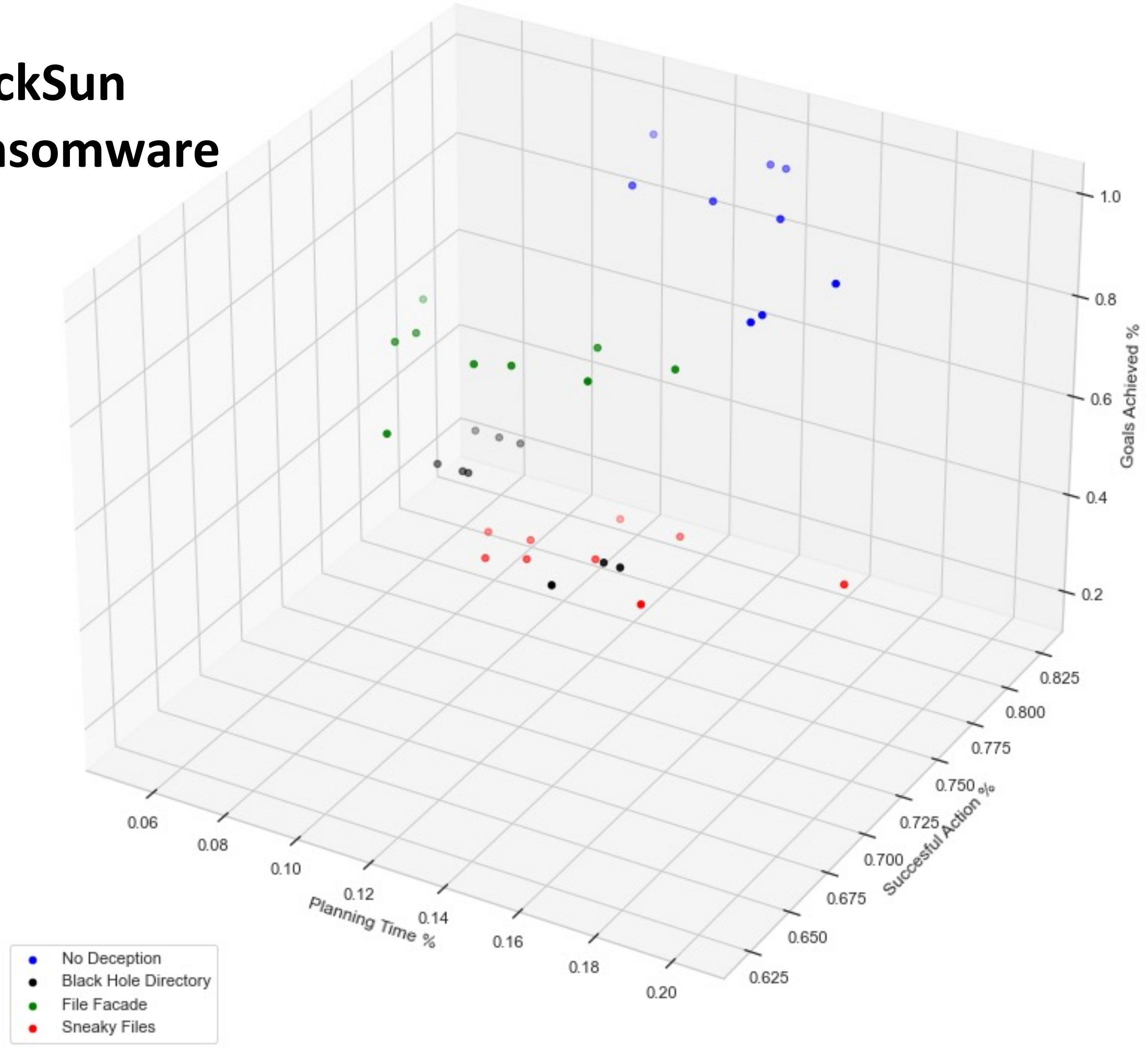
Obviously.

Results: Did the cyber deceptions work?

(Simple) Thief



BlackSun
Ransomware



Results: Did the cyber deceptions work?

General

- Cyber deceptions had a clear (negative) performance effect on the cyber planners, across all adversaries.
- The superiority of the advanced planners was really demonstrated with the BlackSun ransomware adversary. (which was the more complex and realistic adversary)

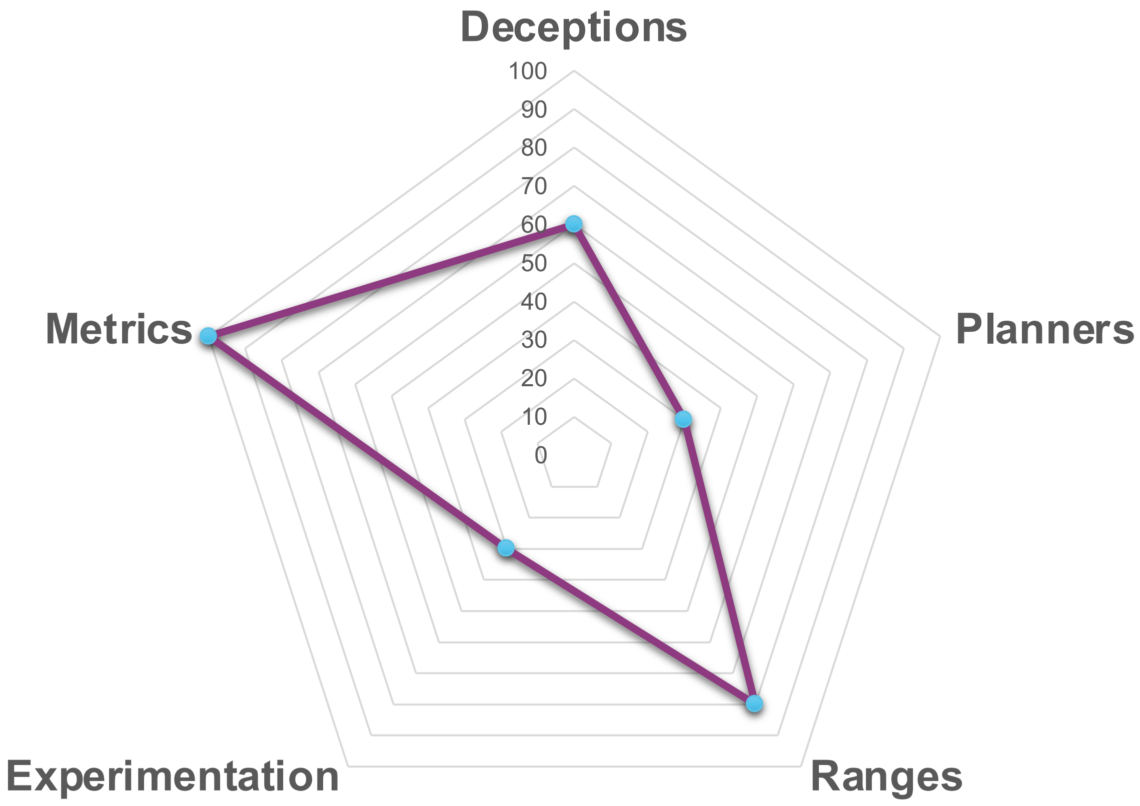
Specific to planner implementations

- **Thief adversary** – advanced planners were faster, but deceptions caused many more failed actions.
- **File Façade deception** – advanced planners had to consider more information which caused significant additional planning time.
- **Black Hole deception** – preventing BlackSun ransomware from any lateral-movement.

Results: Efficacy of the Mirage system

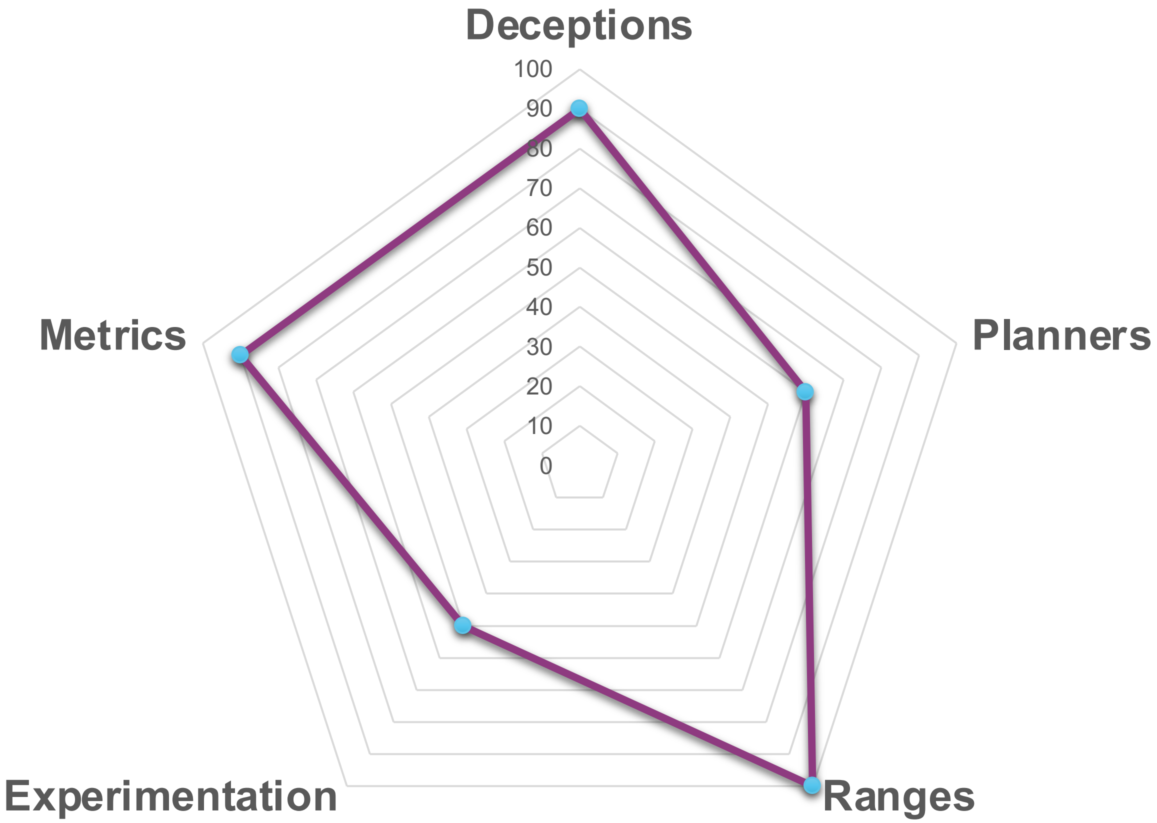
Modularity & Scalability

How hard is it to create and test more of each component?



Practicality

How realistic is each component?



What's next for Mirage?

- ❖ Simulation
- ❖ Cyber gyms for experimentation
- ❖ High fidelity cyber environments for deception simulation
- ❖ Target capabilities:
  - Machine-speed offensive cyber simulations
  - Easy, programmatic defining of cyber deceptions
  - Large scale experimentation

Under Active Development

black hat USA 2023

Q & A

Acknowledgements

This project would not have been possible without code and technical contributions from Zoe Cheuvront, Ethan Michalak, and David Davila.

Contact

Send compliments and kudos to mkouremetis@mitre.org 😎

Send criticisms and challenges to ralford@mitre.org

This work is funded by MITRE's Independent R&D Program.

MITRE