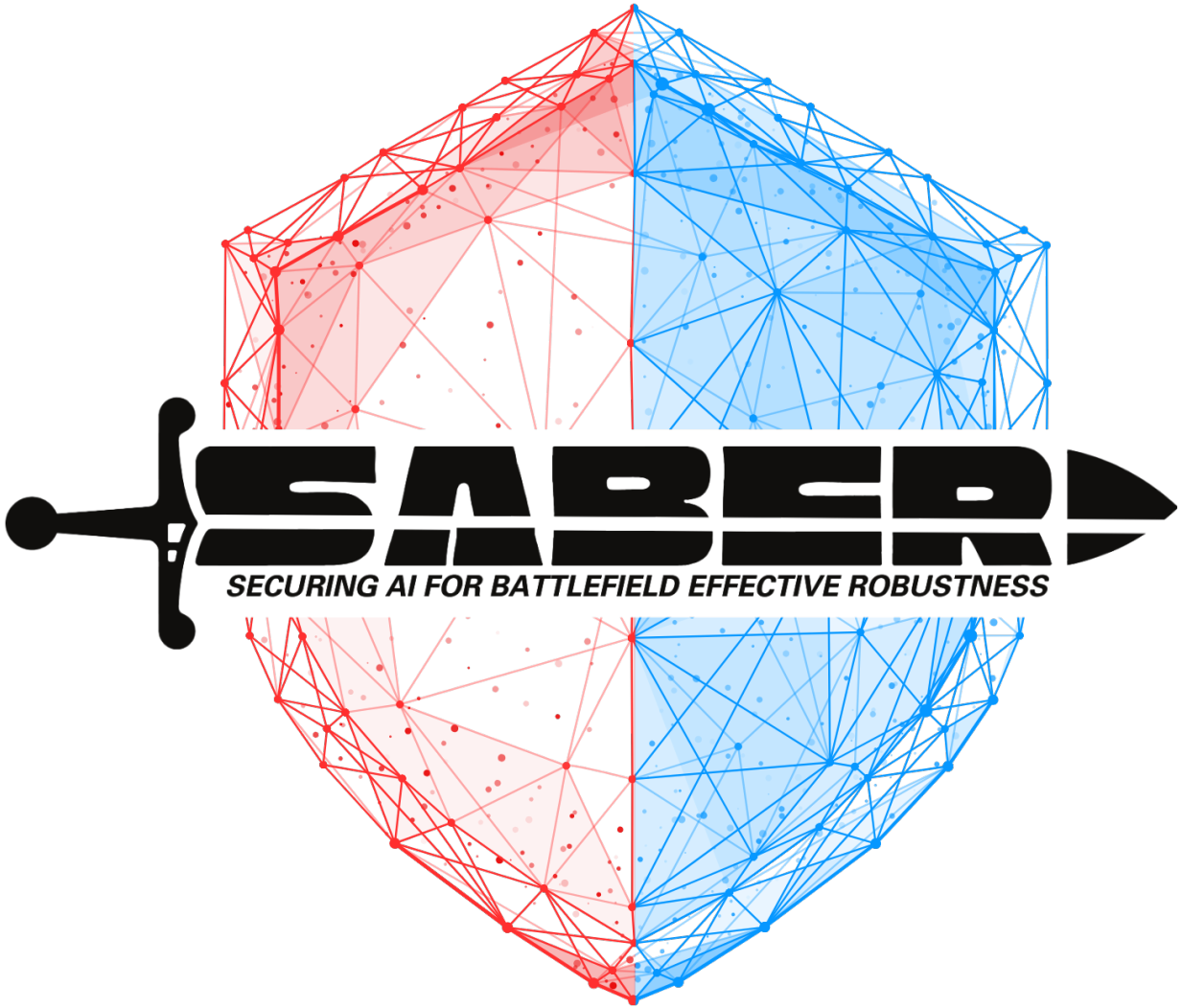


Securing Artificial Intelligence for Battlefield Effective Robustness (SABER)

LTC Nathaniel D. Bastian, PhD
Information Innovation Office (I2O)

Proposers Day
March 12, 2025

Agenda – Open Session

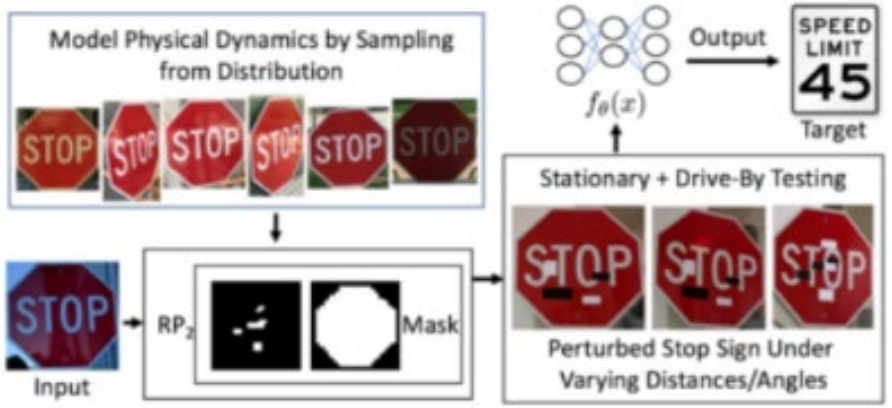
8:00 AM - 9:00 AM	Arrival/Check-In	Executive Conference Center
9:00 AM - 9:05 AM	Opening Remarks	LTC Nathaniel D. Bastian, PhD Program Manager, DARPA I2O
9:05 AM - 9:10 AM	Security Review	Mr. Justin Winburn Program Secuirty Representative
9:10 AM - 9:30 AM	DARPA CMO Review	Contracting Officer Representative
9:30 AM - 10:15 AM	DARPA PM Presentation	LTC Nathaniel D. Bastian, PhD Program Manager, DARPA I2O
10:15 AM - 10:30 AM	Break / Attendees submit to Q&A	
10:30 AM - 11:30 AM	Informal Teaming / PM reviews Q&A	LTC Nathaniel D. Bastian, PhD Program Manager, DARPA I2O
11:30 AM -12:00 PM	PM Addresses Q&A	LTC Nathaniel D. Bastian, PhD Program Manager, DARPA I2O

CMO = Contract Management Office
PM = Program Manager
DARPA = Defense Advanced Research Projects Agency

Operational security risks of AI-enabled battlefield systems are unknown

https://www.defenseone.com/technology/2024/10/new-ai-powered-strike-drone-shows-how-quickly-battlefield-autonomy-evolving/400179/?oref=d1-topic-lander-river

AI-enabled battlefield systems can help improve the speed, quality and accuracy of decisions in the field, providing a decisive advantage



K. Eykholt et al., "Robust Physical-World attacks on deep learning models," arXiv.org, Jul. 27, 2017. https://arxiv.org/abs/1707.08945

Adversarial AI attacks have not been practically demonstrated in operational settings

Technical Hypothesis:
An AI red team employing SoA PACE techniques and red-teaming tools to execute novel AI kill chain TTPs can operationally assess AI-enabled battlefield systems more effectively than a team without them

Disruption Target:
Build an exemplar AI red team that can continuously integrate and employ emerging counter-AI techniques and tools, establishing a sustainable model for an operational AI red-teaming process

AI = Artificial Intelligence
SoA = State-of-the-Art
PACE = Physical+Adversarial AI+Cyber+Electronic Warfare
TTP = Tactics, Techniques and Procedures

AI-enabled battlefield systems are used without knowing the risks

Source: https://www.nytimes.com/2024/07/02/technology/ukraine-war-ai-weapons.html

Drones

Live video feed taken from an AI-enabled drone, identifying an asset…

… then autonomously delivering its firing effect

Turrets

AI-enabled auto-turrets are giving operators the ability to work remotely with enhanced targeting…

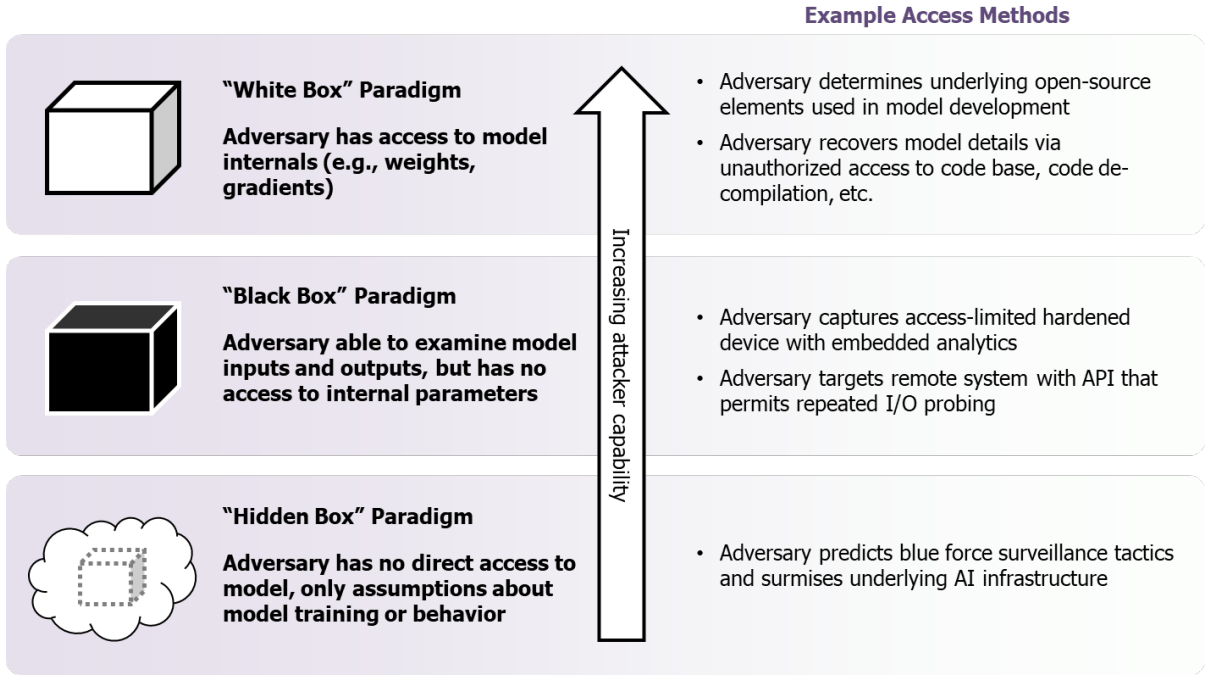
… and can be mounted on autonomous ground robots, relying on auto aim to identify and eliminate targets

Operational AI red-teaming does not exist, making operationally impactful vulnerabilities unknown

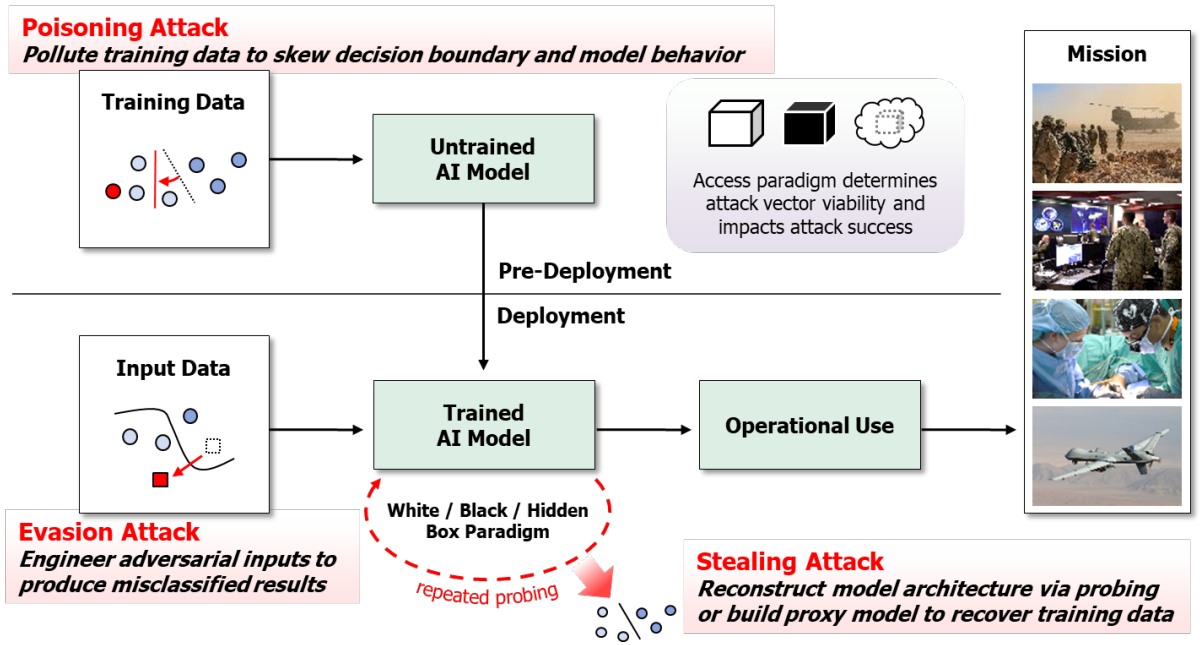
DoD = Department of Defense
AI = Artificial Intelligence

Operational AI red-teaming needs PACE techniques to execute AI kill chain

Adversarial AI access paradigms



Adversarial AI attack categories

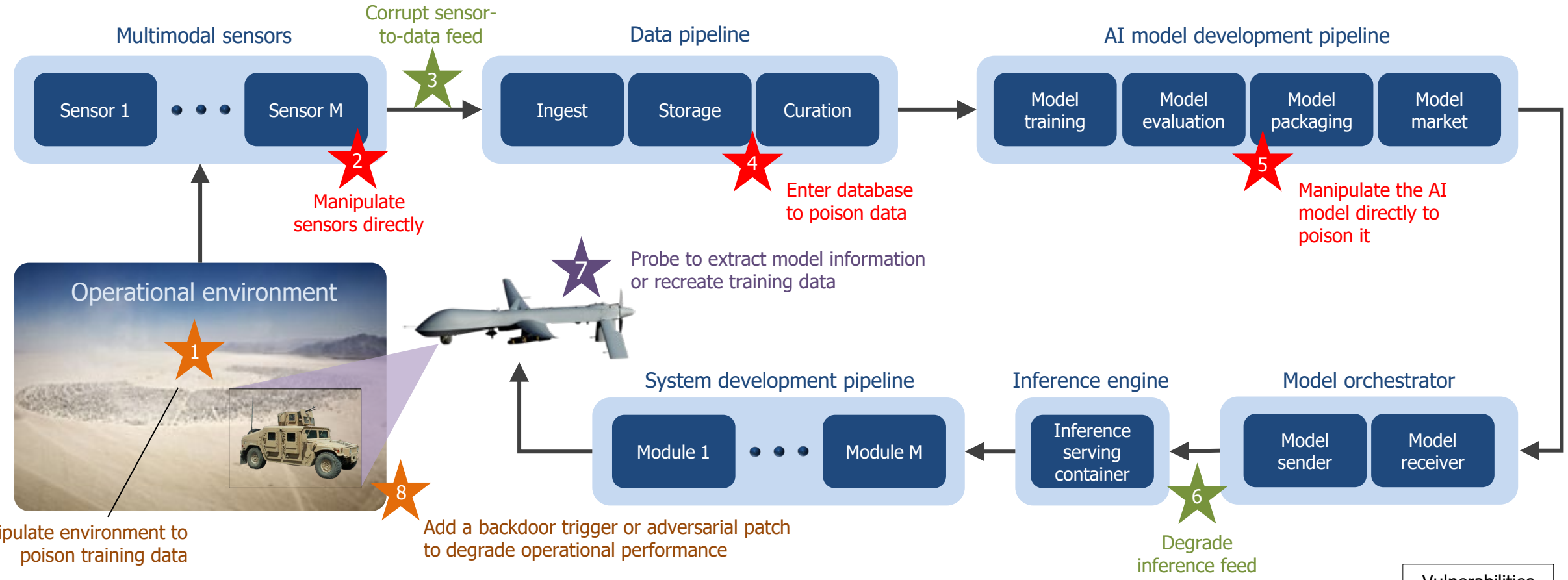


No work has coupled physical, adversarial AI, cyber and EW attacks to create AI kill chain TTPs needed for operational AI red-teaming

Most previous academic work ignores physically realizable attacks and the full-system pipeline

AI = Artificial Intelligence
EW = Electronic Warfare
PACE = Physical+Adversarial AI+Cyber+EW
TTP = Tactics, Techniques and Procedures

Example AI-enabled battlefield system pipeline for vulnerability discovery

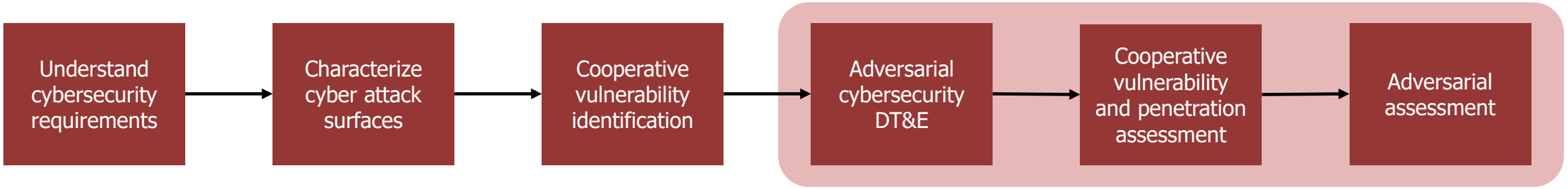


Operational AI red teaming requires the creation and execution of novel AI kill chain TTPs that effectively employs SoA PACE techniques and red-teaming tools

AI = Artificial Intelligence
EW = Electronic Warfare
TTPs = Tactics, Techniques and Procedures
SoA = State-of-the-Art
PACE = Physical+Adversarial AI+Cyber+EW

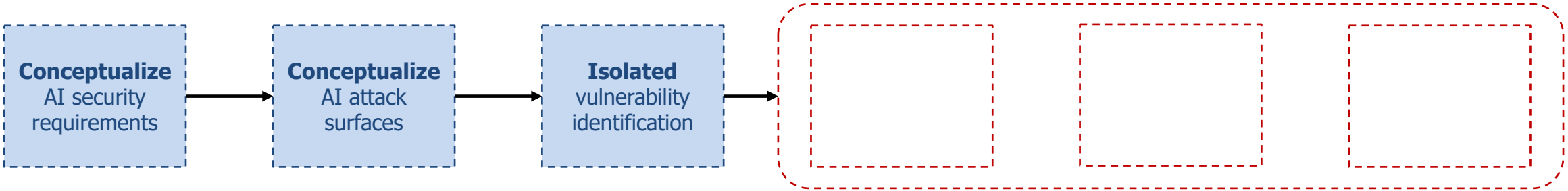
Today: Assessment of AI-enabled battlefield systems does not exist

DoD cybersecurity operational test and evaluation [1]



Cyber red-teaming phases

DoD AI security operational test and evaluation



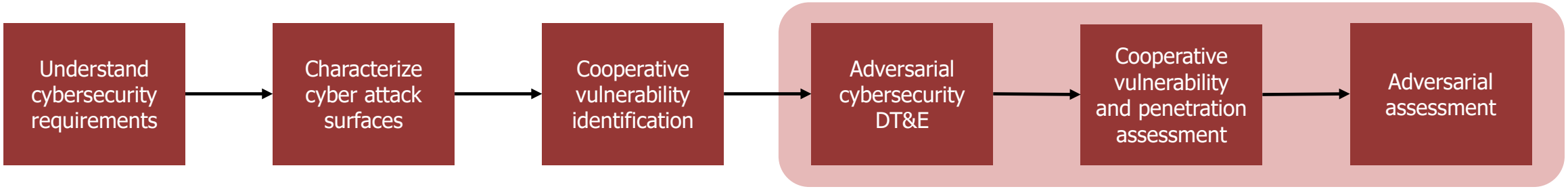
AI security red-teaming phases does not exist

DoD = Department of Defense
DT&E = Developmental Test & Evaluation
AI = Artificial Intelligence

[1] DoD Cybersecurity Test and Evaluation Guidebook, Version 2.0, Change 1. February 2020.

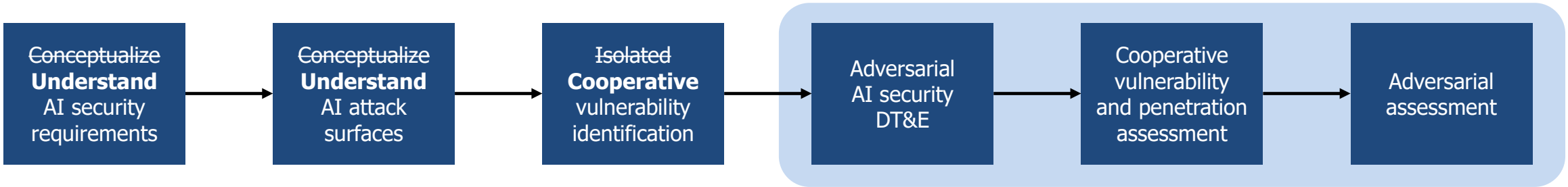
Tomorrow: Sustainable model for operational AI red-teaming capability

DoD cybersecurity operational test and evaluation [1]



Cyber red-teaming phases

DoD AI security operational test and evaluation



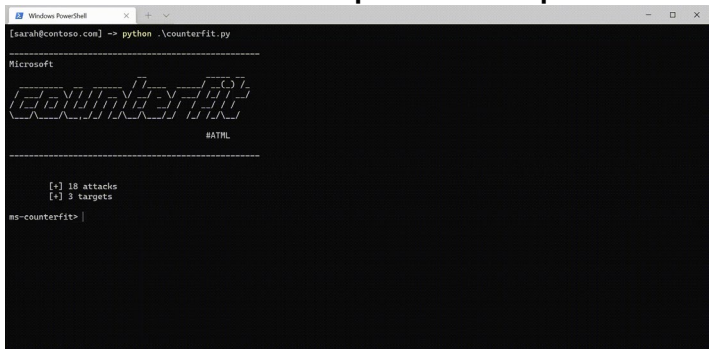
An AI red team employing SoA PACE techniques and red-teaming tools to execute novel AI kill chain TTPs

DoD = Department of Defense
AI = Artificial Intelligence
SoA = State-of-the-Art
PACE = Physical+Adversarial AI+Cyber+Electronic Warfare
TTP = Tactics, Techniques and Procedures

[1] DoD Cybersecurity Test and Evaluation Guidebook, Version 2.0, Change 1. February 2020.

Major gaps in the AI red-teaming ecosystem that DARPA can enable

No adequate techniques and tools exist for future DoD AI red teams to perform operational AI red-teaming



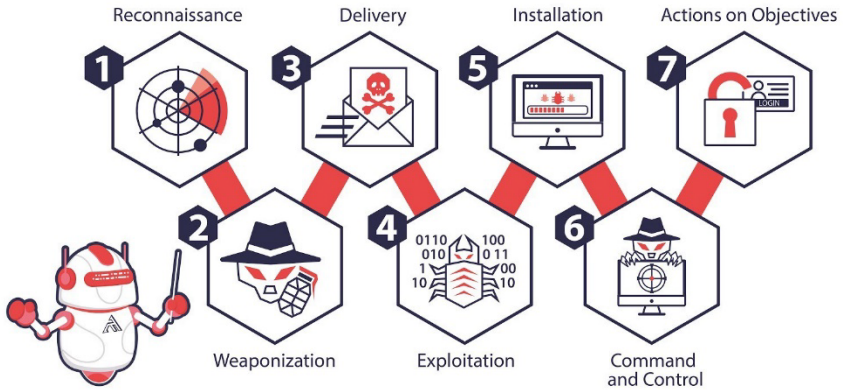
Adversarial Robustness Toolbox

(left): https://www.microsoft.com/en-us/security/blog/2021/05/03/ai-security-risk-assessment-using-counterfit/
(right): https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

DoD has no testing ground(s) for future DoD AI red teams to perform cyber-physical-digital AI red-teaming

(left): https://www.army.mil/article/243923/tech_demos_to_provide_better_glimpse_of_reality_for_soldiers_scientists
(right): https://southernindefense.com/portfolio-items/muscatatuck-urban-training-center-mutc/

Technology	**Places**
Processes	**People**



https://pentestgurus.com/kill-chain-pros-cons/

There are no operationally-guided TTPs to create and execute AI kill chains needed for use by future DoD AI red teams

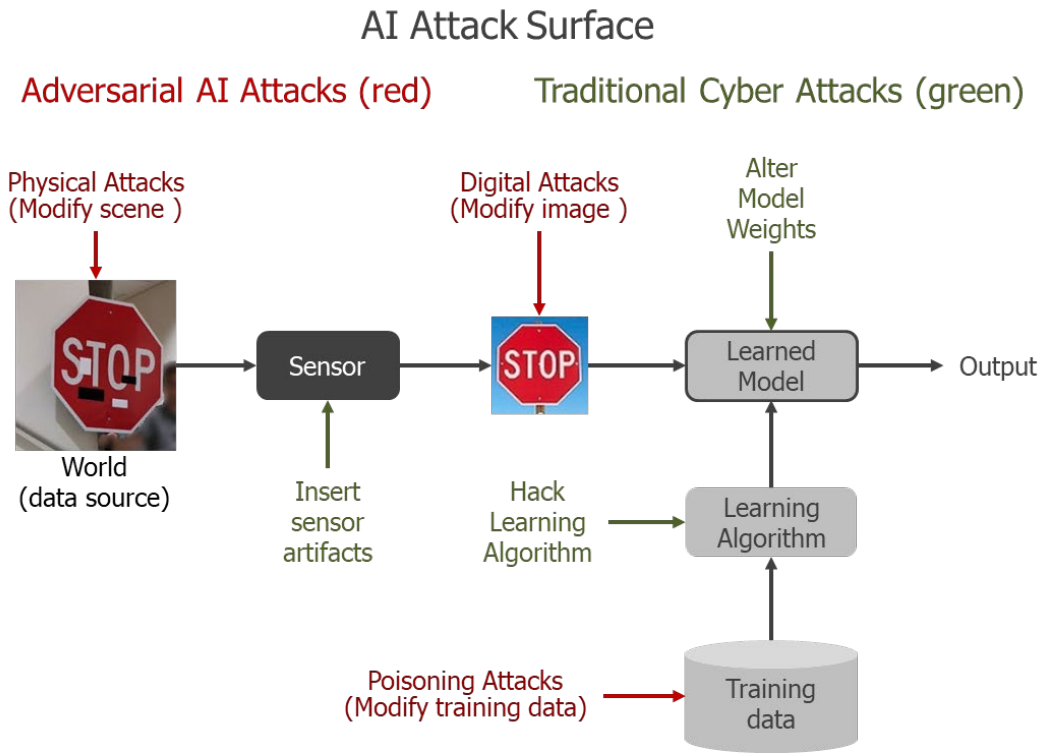
https://simspace.com/blog/what-is-a-red-team-exercise-and-why-is-it-important-simspace/

No DoD organization(s) has proper guidelines to certify, accredit and employ future DoD AI red teams

DoD = Department of Defense
AI = Artificial Intelligence
TTPs = Tactics, Techniques, and Procedures

AI attack effects driving program design for an operational AI red team

AI red teams must consider the novel class of intersectional AI and cyber/EW attack vectors for AI attack effects



Attack Effect \ Attack Vector	Cyber/EW	AI
AI	Cyber/EW attacks on AI-enabled systems (in scope)	AI-informed attacks on AI-enabled systems (in scope)
Cyber/EW	Traditional cyber/EW attacks (not in scope)	AI-assisted cyber/EW attacks (not in scope)

Cyber/EW attack effects are NOT in scope

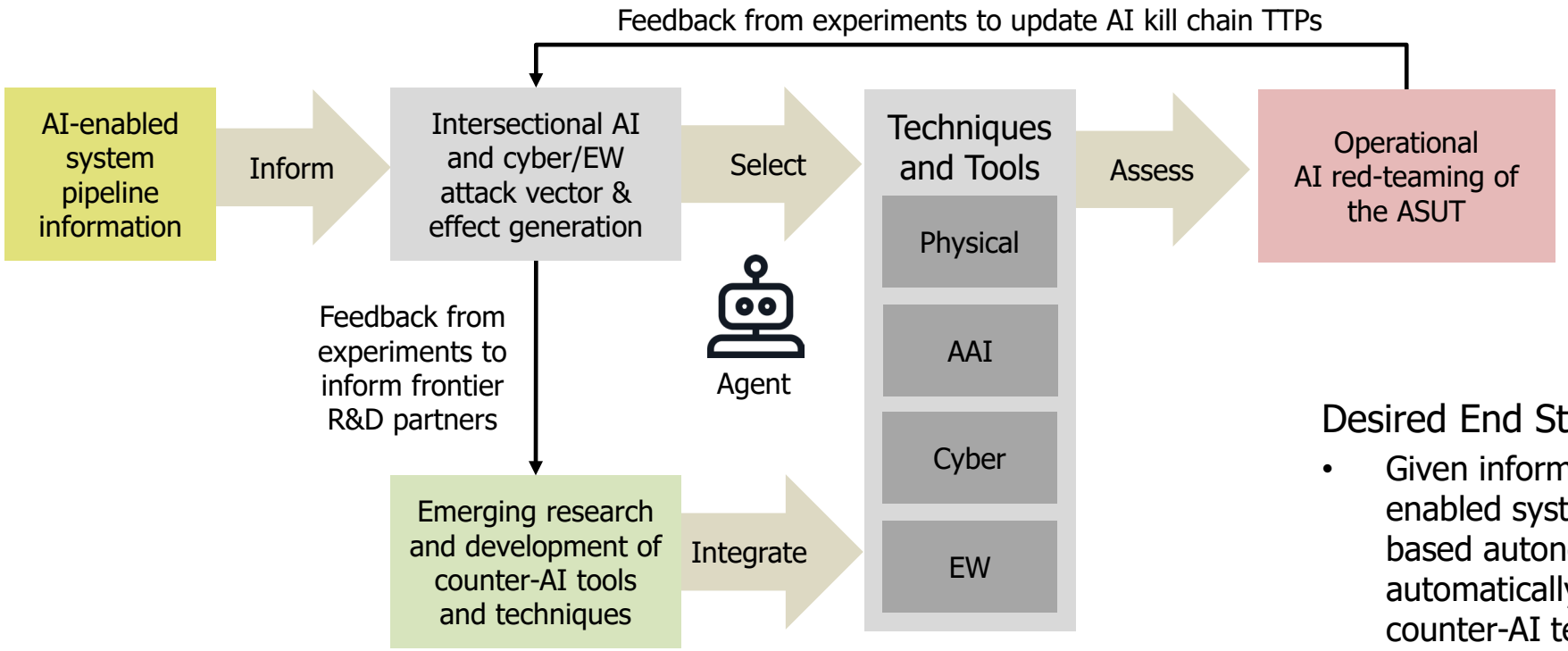
AI attack effects are in scope

Mature
Adolescent
Immature

AI = Artificial Intelligence
EW = Electronic Warfare

A sustainable model for an operational AI red-teaming process

Hypothesis: An AI red team employing SoA PACE techniques and red teaming tools to execute novel AI kill chain TTPs can operationally assess AI-enabled battlefield systems more effectively than a team without them



An agent prescribes which techniques and tools to employ rather than randomly selecting vectors/effects

Desired End State

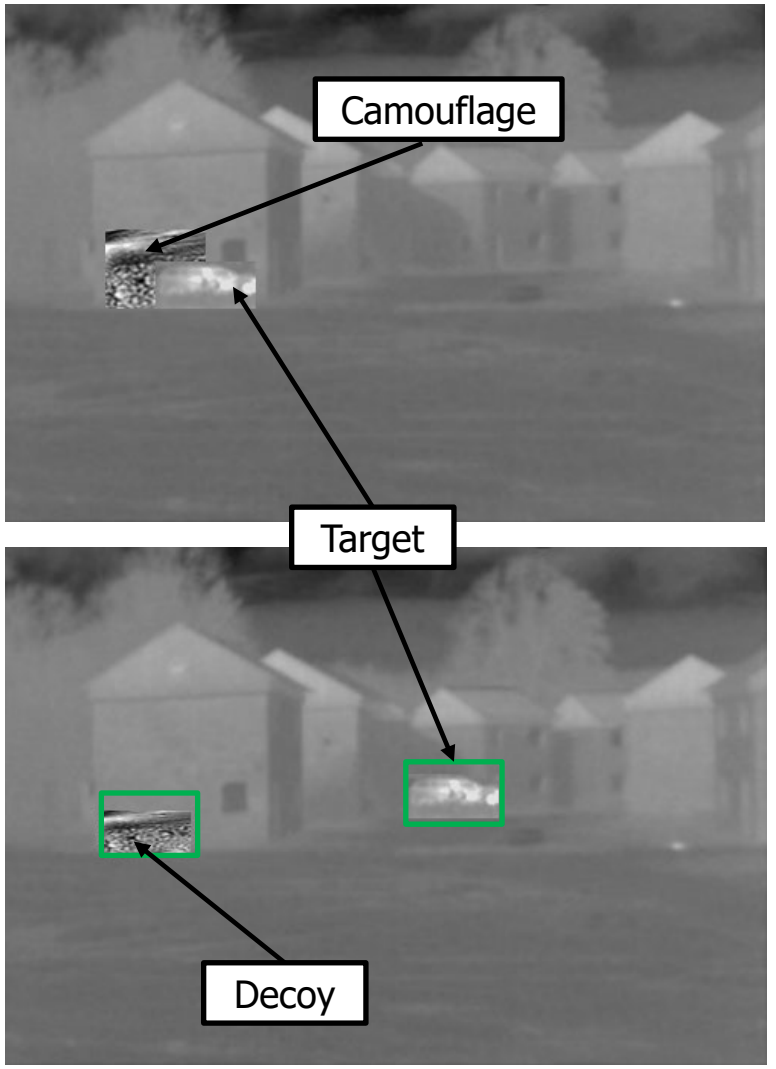
- Given information about the AI-enabled system pipeline, an agent-based autonomous red-teaming tool automatically recommends effective counter-AI techniques/tools to employ to operationally assess the ASUT

AI = Artificial Intelligence
AAI = Adversarial AI
EW = Electronic Warfare
ASUT = AI-enabled System-under-Test
TTP = Tactics, Techniques and Procedures

Program design: execute experimentation-driven AI security exercises

SABER-OpX: A series of high-fidelity operational exercises to iteratively assess (red-team) already developed AI-enabled systems in the settings where they will be deployed

- Develop a "model" for an operational AI red-teaming process
- Evaluate the discriminative AI paradigm to scale across "battlefield" systems
- Emphasize AI/cyber/EW attack vectors that delivers AI attack effects
- Develop operationally-guided TTPs to create and execute AI kill chains
- Continuously integrate emerging techniques/tools into the AI red team
- Estimate operational security risk of deployed AI-enabled systems
- Determine and investigate the trade-space of AI red team effectiveness



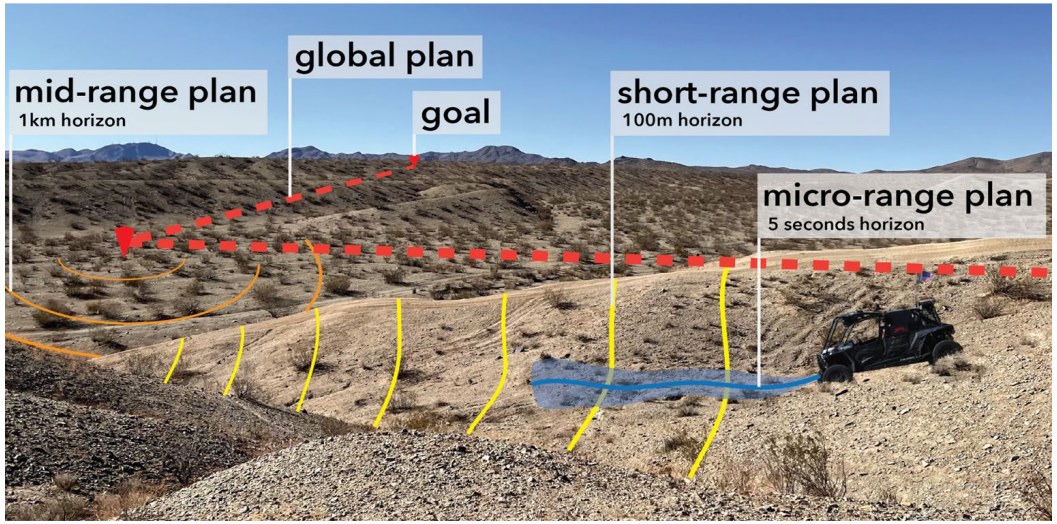
https://www.bbc.co.uk/wiltshire/content/articles/2007/12/31/imber_2007_feature.shtml

Notional physically realizable adversarial attacks on infrared AI-enabled aided target recognition system

AI = Artificial Intelligence
EW = Elecronic Warfare
TTP = Tactics, Techniques and Procedures

SABER-OpX #1: AI-enabled Autonomous Ground Vehicle

Blue team goal:
Enable autonomous UGV to navigate terrain to a goal



- Experiments evaluate AI-enabled autonomy in increasingly difficult operating environments
- Main metric: Increase average autonomous speed

AI red team goal:
Disable autonomous UGV to navigate terrain to a goal

Sensor Data	Perception	Planning	Low Level control
	To Directly Effect:		To Indirectly Effect:
Perturb data using techniques and tools…	• Localization • Semantic classification • Voxel mapping • World modeling	• Traversability • Uncertainty estimation • Multi-range planners	• Vehicle controls • Overall speed • Overall vehicle safety • Vehicle navigation

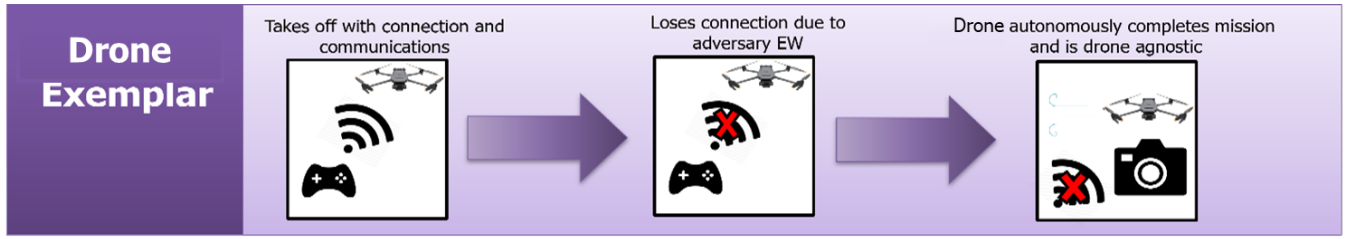
AI components

- AI red team will manipulate the autonomy by using PACE techniques/tools to degrade the AI-based perception
- Main metric: Reduce average autonomous speed

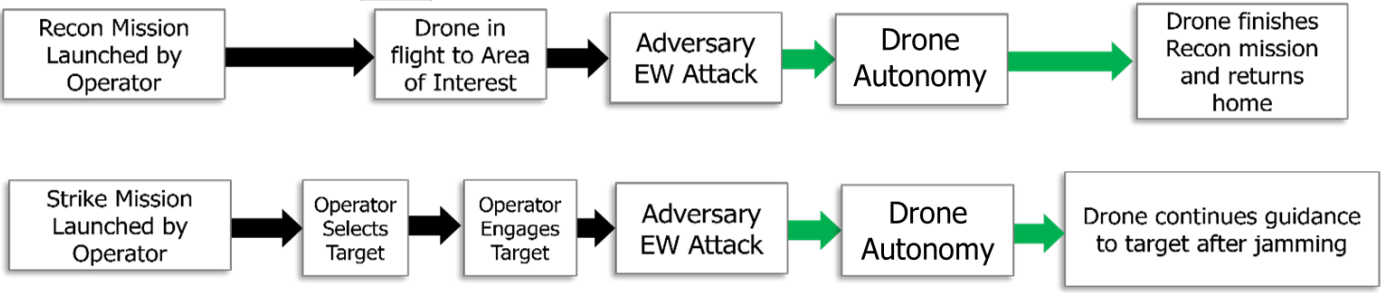
PACE = Physical+Adversarial AI+Cyber+Electronic Warfare
UGV = Unmanned Ground Vehicle
AI = Artificial Intelligence

SABER-OpX #2: AI-enabled Autonomous Drone (Rotary-Wing)

Blue team goal:
Develop drone autonomy to harden system against EW

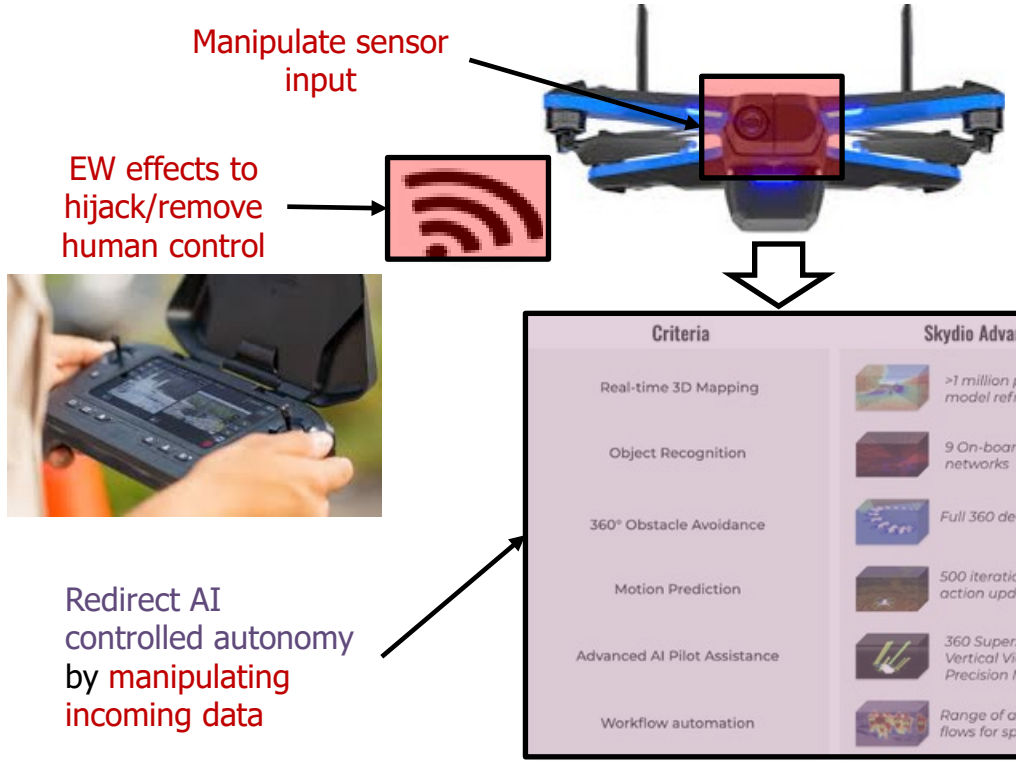


Examples of how we will effect the outcome of a pre-selected mission from an Operator



- Ongoing work to rapidly harden drone platforms to counter adversary EW effects, integrating AI to enable a drone to carry out its mission without direct human control
- Main metric: Success rate of autonomy to complete objective (recon mission, strike mission)

AI red team goal:
Degrade autonomy to reduce drone system effectiveness



(left): https://www.skydio.com/x10
(right top): https://www.skydio.com/skydio-2-plus-enterprise
(right bottom): https://medium.com/skydio/skydio-autonomy-a-new-age-of-drone-intelligence-12346111b2f1

- AI red team will manipulate the autonomy by using PACE techniques/tools to degrade the AI-based perception
- Main metric: Failure rate of autonomy to complete objective (recon mission, strike mission)

AI = Artificial Intelligence
EW = Electronic Warfare
PACE = Physical+Adversarial AI+Cyber+EW

SABER-OpX uses a DOE approach to evaluate discriminative AI

The ASUTs represent variety among the operational characteristics that impact counter-AI effects in near-term discriminative AI, implying that AI kill chain TTPs over the ASUTs should be applicable to new battlefield systems

Operational Characteristics	SABER-OpX #1	SABER-OpX #2	Joint Coverage?
View Perspective	Ground	Air	✓
Illumination	Ambient+Self	Ambient Only	✓
Sensor Modalities	Optical, Thermal, GPS, Lidar, RADAR, NDVI	Optical, optional thermal	✓
Motion Dynamics	2D, constrained	3D, unconstrained	✓
Effect Delivery	Predictable Paths	Omnidirectional approach window	✓

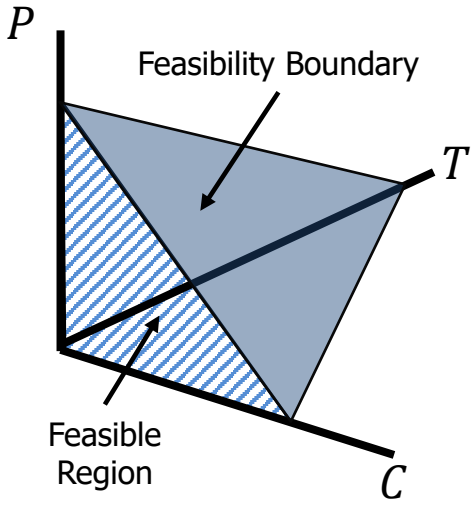
DOE = Design of Experiments
AI = Artificial Intelligence
ASUT = AI-enabled System-under-Test
TTP = Techniques, Tactics, and Procedures
NDVI = Normalized Difference Vegetation Index

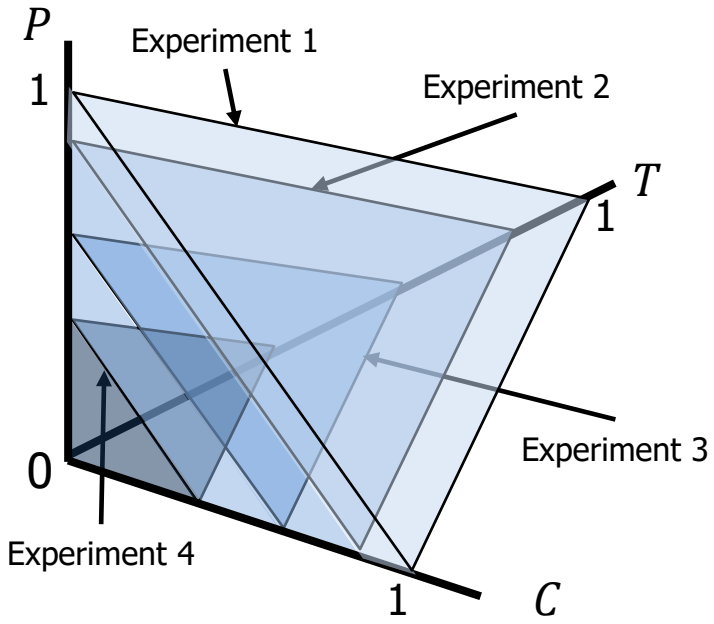
SABER program metric

Core program metric: AI red team effectiveness

$$R = 1/(\theta_P P + \theta_T T + \theta_C C)$$

- P – Performance degradation of Blue system given AI attack effect
  - $P = \frac{N}{B}$, see chart per SABER-OpX
- T – Time (normalized) to generate the AI attack effect
- C – Cost (normalized) to generate the AI attack effect
- θ_i – Importance given to each criteria (sums to one)





$1/R$ is a proxy measure for operational security risk

Each SABER-OpX will be a 9-month sprint, with experiments occurring at the end of months 1, 3, 6, and 9

For each SABER-OpX, we aim to improve R:
- Experiment 1: establish baseline, R_{base}
- Experiment 2: 10% increase above R_{base} (goal)
- Experiment 3: 20% increase above R_{base} (metric)
- Experiment 4: 40% increase above R_{base} (goal)

SABER-OpX experiment metrics

SABER-OpX	System Metric (B)	Negative System Metric (N)
OpX #1	Maximize average autonomy speed	Minimize average autonomy speed
OpX #2	Mission success rate	Mission failure rate

B – Measure of performance of the Blue system *without* red degradation
N – Measure of performance of the Blue system *with* red degradation

AI = Artificial Intelligence

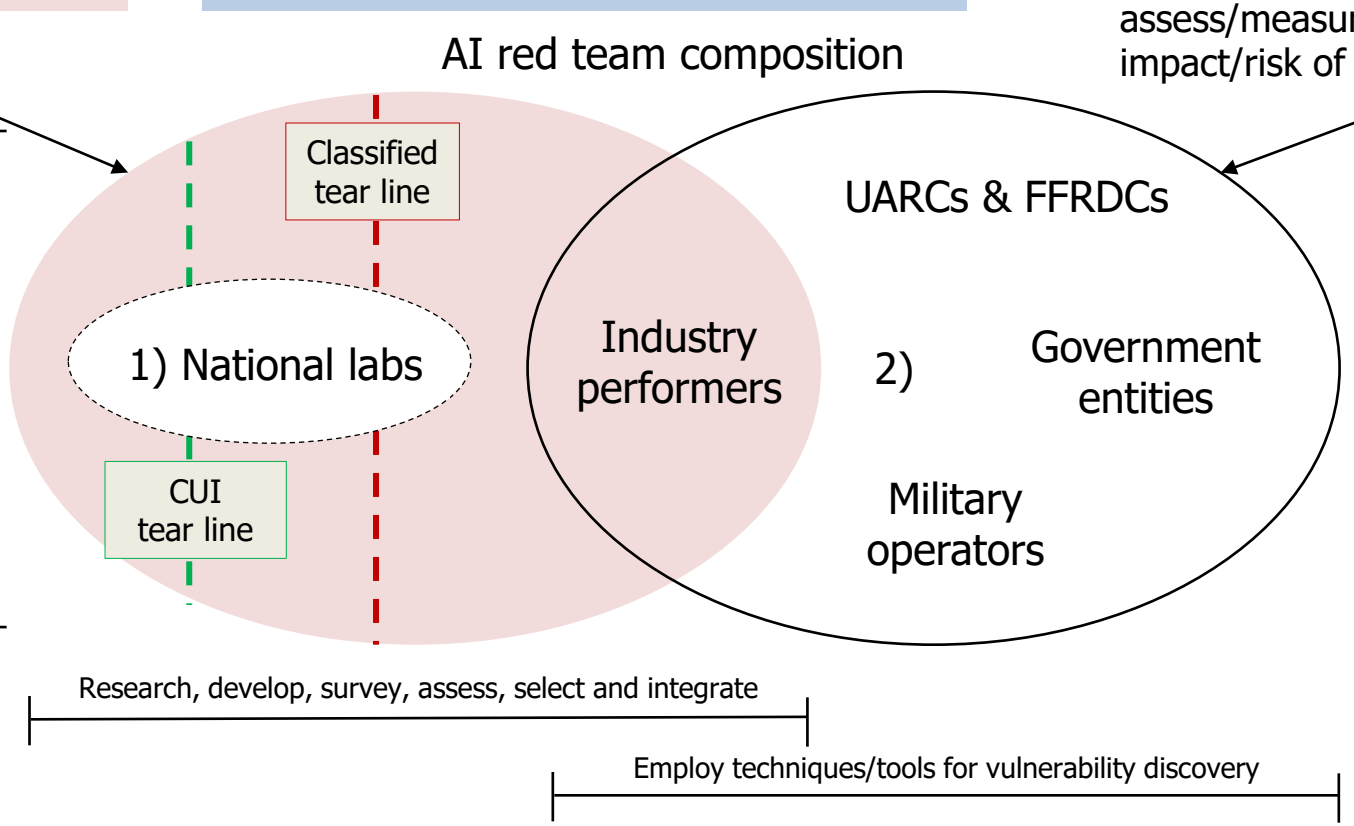
SABER program structure

TT1 – Red Team
- Survey, assess, select, integrate, and employ techniques and tools to red-team Blue Teams' AI-enabled systems

TT2 – Blue Teams
- Provide access and use of their existing AI-enabled battlefield systems and support experimentation for operational assessment by the AI red team

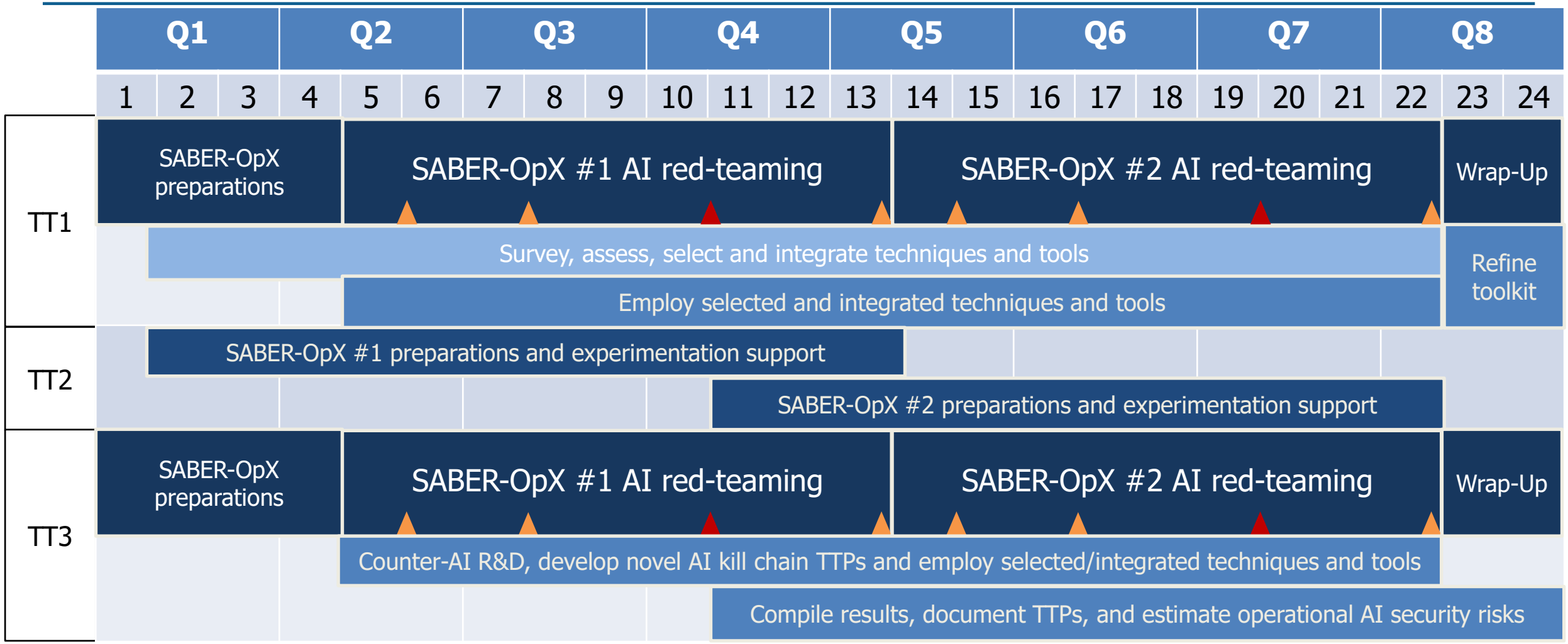
TT3 – Government Team
- 1) Research and develop novel PACE techniques and red-teaming tools
- 2) Develop novel AI kill chain TTPs and employ selected and integrated PACE techniques and red-teaming tools; assess/measure the operational impact/risk of adversarial attacks

Foster continuous integration of ever-emerging counter-AI techniques and tools from research and development partners into the AI red team for iterative selection/employment



PACE = Physical+Adversarial AI+Cyber+Electronic Warfare
AI = Artificial Intelligence
TTPs = Tactics, Techniques and Procedures
UARC = University Affiliated Research Center
FFRDC = Federally Funded Research and Development Center

SABER program schedule



▲ = SABER OpX Experiment

▲ = SABER OpX Metric Evaluation

AI = Artificial Intelligence
TTPs = Tactical, Techniques, Procedures

www.darpa.mil