AI Red Teaming

Joseph Lucas, AI Security Researcher | Assurance and Security for AI-enabled Systems, April 2024

"AI Red Teaming" has come to mean different things to different organizations. All those definitions are wrong, but some are useful. Here's ours.

AI Red Teaming

BLUF

An AI Red Team is a **friendly force** that provides **targeted stimulus** to determine the **effectiveness of security controls** for **AI-integrated systems** and provides **actionable recommendations** to reduce the likelihood or effectiveness of adversarial action.

AI Red Teaming

Why [it matters]?

- Why red team?
  - Measure system reaction to *known unknown* and identifies *unknown unknown*
  - Objective-guided adversarial stimulus
  - Risk identification and calibration
- Red teaming provides system operators with data about how their system performs under adversarial conditions

AI Red Teaming

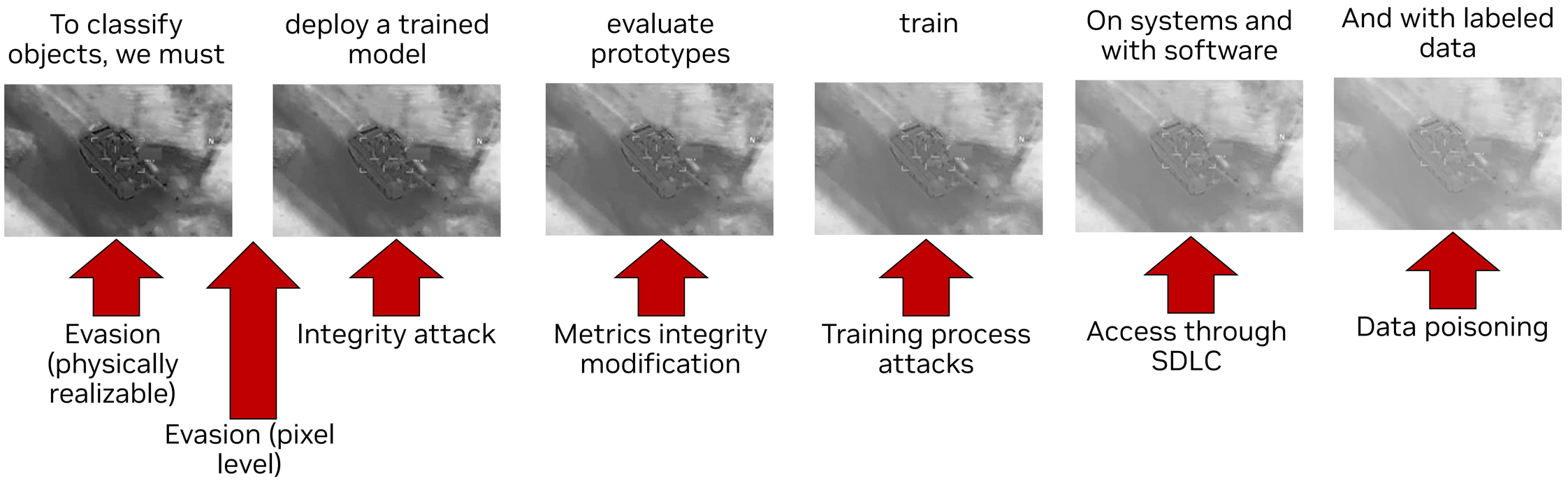
Why [it might be unconventional]?

- In some contexts, we red team a "final product" as part of a final T&E:
  - BCTs going to CTC
  - CMTs before being FOC
- The final performance of AI systems are highly dependent on inputs at several points during the lifecycle:
  - Test time
  - Training time
  - Prototyping
- Comprehensive test and evaluation against the complete range of inputs is difficult to do and interpret.
  - And we know there are small epsilon attacks
- “Final product” T&E is insufficient for AI-integrated systems.

AI Red Teaming

MQ9

- "The system has 368 cameras capable of capturing five million pixels each to create an image of about 1.8 billion pixels" - General Atomics MQ-9 Reaper - Wikipedia
  - 42k x 42k
  - Monochrome – Each pixel is [0, 255]
  - State space? Huge.

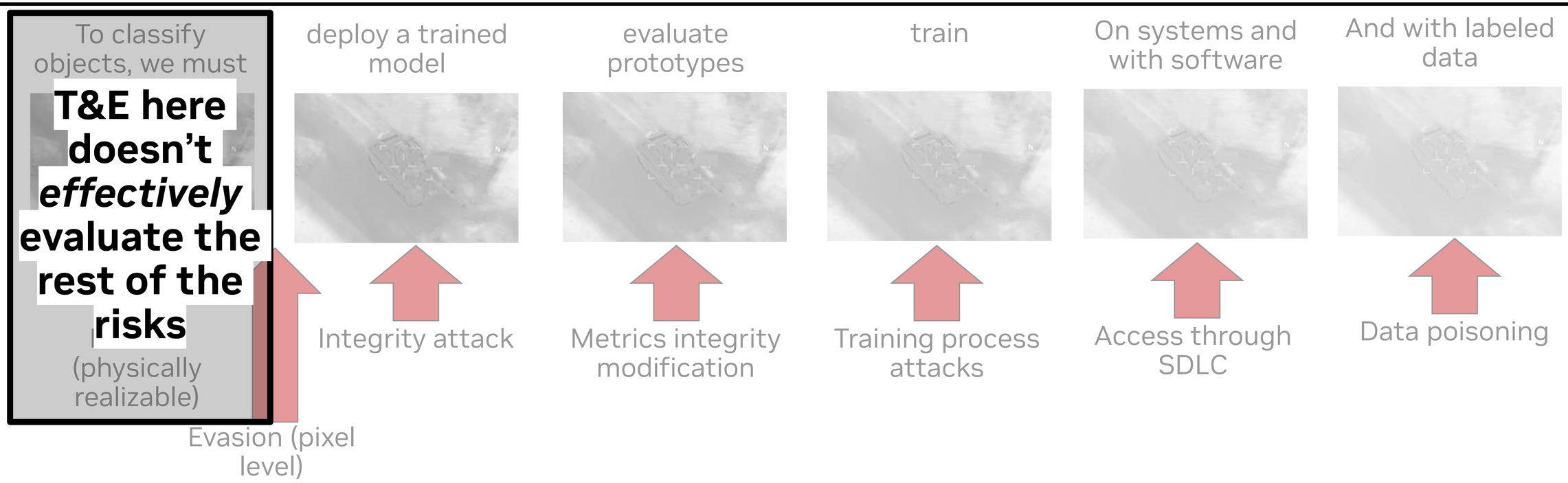


US MQ-9 Reaper Drone Takes Out Russian T-72 Tank in Syria | Military.com

AI Red Teaming

MQ9

- "The system has 368 cameras capable of capturing five million pixels each to create an image of about 1.8 billion pixels" - General Atomics MQ-9 Reaper - Wikipedia
  - 42k x 42k
  - Monochrome – Each pixel is [0, 255]
  - State space? Huge.



US MQ-9 Reaper Drone Takes Out Russian T-72 Tank in Syria | Military.com

AI Red Teaming

How [we got here]?

- We want to deploy AI in security-critical contexts.
- Performance must be measured under adversarial conditions
- **Adversarial**: conditions and stimulus most disadvantageous to friendly objectives
- We don't always know *who* the adversary is, what *their objectives* are, or what *capabilities* they have
  - Sometimes intelligence-driven adversarial emulation
  - But sometimes systems may have to operate against unknown unknowns

AI Red Teaming

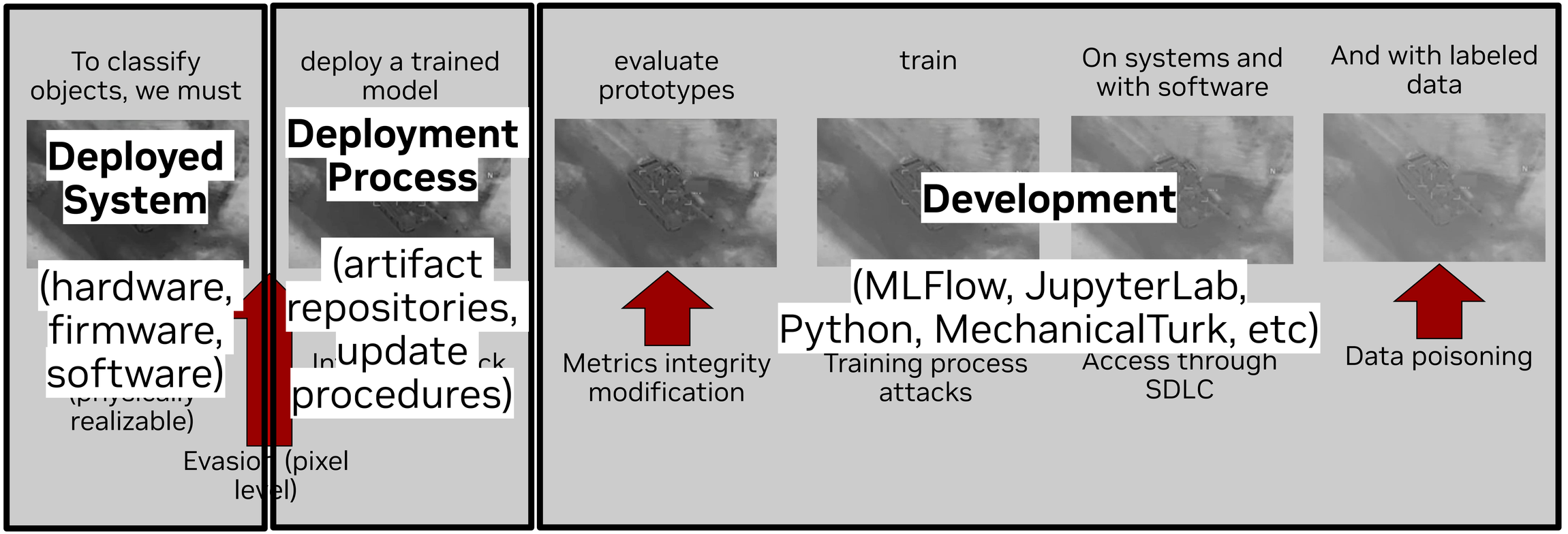
How [***not*** to do it]?

- Red Teams improve the security of the system.
  - No stunt hacking.
- Is the target a Model or a System? (it's a system)
- Security is a system property
  - The most secure operating system can be "hacked" with bad passwords
- And the security of the deployed system relies on the security of all inputs
  - Inputs are not just raw data
    - How much do you know about computational photography?
- "The system has 368 cameras capable of capturing five million pixels each to create an image of about 1.8 billion pixels" - General Atomics MQ-9 Reaper – Wikipedia
  - Image processing -> 368 images are probably not just tiled together. They're transformed (rotation, stretching, error correction, etc). **Those transformations are inputs.** They're not part of the model, but they're an important security component of the system. (not to mention sensor fusion)

AI Red Teaming

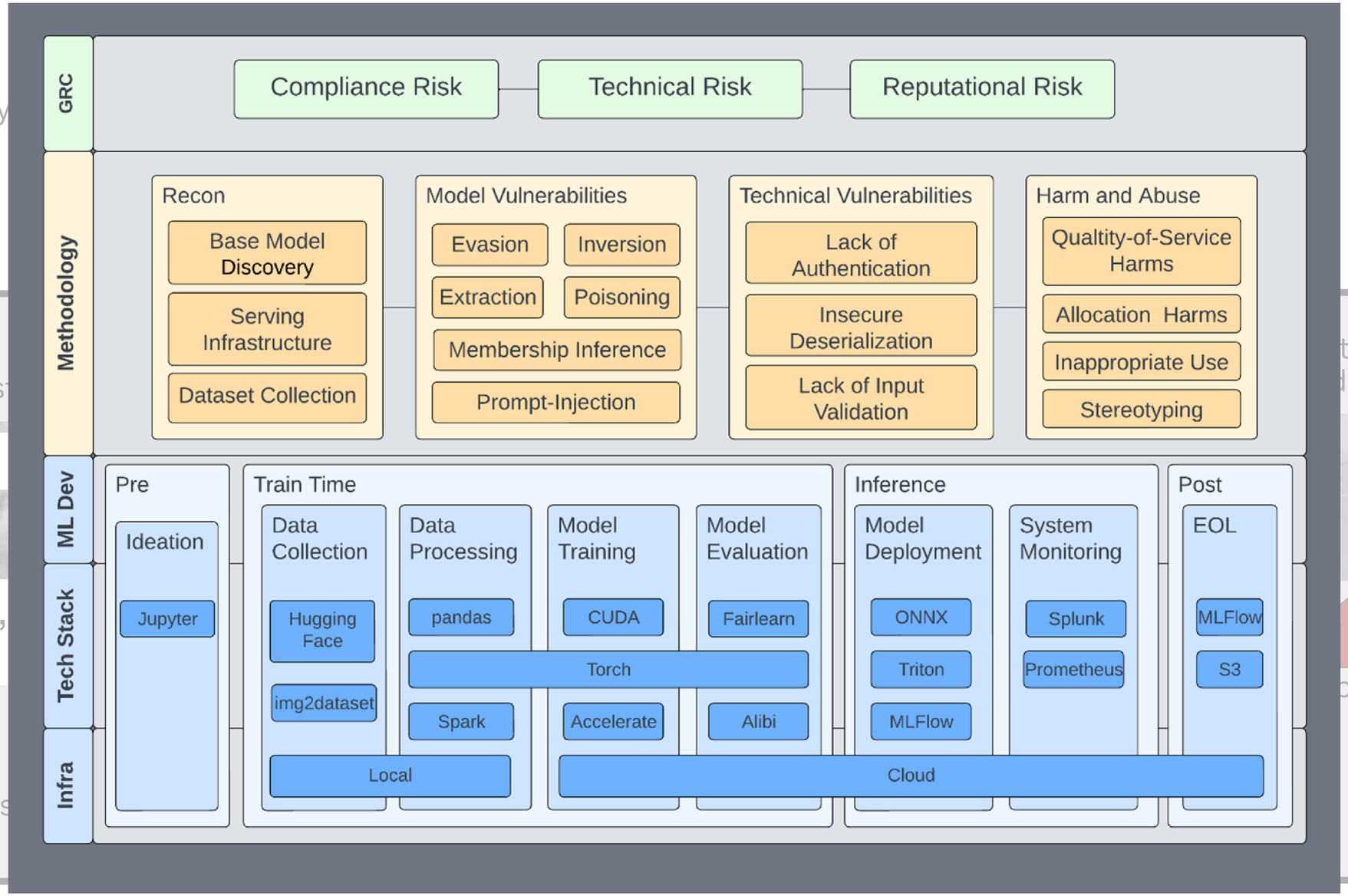
How [and when to do it]?

- Deployed systems *can* be targets, but they're not the *only* targets



AI Red Teaming

How [and when to do it]?



AI Red Teaming

How [to build the capability]?

- What _____ does an AI Red Teamer need?
  - Skills
  - Tools
  - Access
  - Partners
- Red Teams aren't _____
  - and those things **are important**
- Red Team(ers) must be:
  - Creative
  - Calibrated
  - Precise

AI Red Teaming

How [to build the capability]?

- **What _____ does an AI Red Teamer need?**
  - **Skills**
  - Tools
  - Access
  - Partners
- Red Teams aren't _____
  - and those things are important
- Red Team(ers) must be:
  - Creative
  - Calibrated
  - Precise

Skills

- Understanding of ML Development Lifecycle
- Ability to develop and use gradient-based ML techniques
- Knowledge of optimization paradigms
- Technical proficiency in the target domains/languages

```
Optimize bit-level mask and perturbations over
specific PE header fields and inject the
resulting transformation into a python venv to
run during training data ETL
```

AI Red Teaming

How [to build the capability]?

- **What _____ does an AI Red Teamer need?**
  - Skills
  - **Tools**
  - Access
  - Partners
- Red Teams aren't _____
  - and those things are important
- Red Team(ers) must be:
  - Creative
  - Calibrated
  - Precise

Tools

- https://github.com/Trusted-AI/adversarial-robustness-toolbox
- https://github.com/QData/TextAttack
- https://github.com/leondz/garak
- https://github.com/JosephTLucas/vger
- **... more all the time**

AI Red Teaming

How [to build the capability]?

- **What _____ does an AI Red Teamer need?**
  - Skills
  - Tools
  - **Access**
  - Partners
- Red Teams aren’t _____
  - and those things are important
- Red Team(ers) must be:
  - Creative
  - Calibrated
  - Precise

Access

- **Network**
- **Host**
- **Cloud Service API**

White card access to improve telemetry and efficiency

AI Red Teaming

How [to build the capability]?

- **What _____ does an AI Red Teamer need?**
  - Skills
  - Tools
  - Access
  - **Partners**
- Red Teams aren’t _____
  - and those things are important
- Red Team(ers) must be:
  - Creative
  - Calibrated
  - Precise

Partners

- Threat Intel -> Adversary emulation
- Traditional host/network access teams

AI Red Teaming

How [to build the capability]?

- What _____ does an AI Red Teamer need?
  - Skills
  - Tools
  - Access
  - Partners
- **Red Teams aren't _____**
  - and those things are important
- Red Team(ers) must be:
  - Creative
  - Calibrated
  - Precise

Aren't

- Tactical offensive units
- QA
- Coverage-based T&E
- Security Automation

AI Red Teaming

How [to build the capability]?

- What _____ does an AI Red Teamer need?
  - Skills
  - Tools
  - Access
  - Partners
- Red Teams aren't _____
  - and those things are important
- **Red Team(ers) must be:**
  - **Creative**
  - Calibrated
  - Precise

Creative

- All models are wrong, but some are useful [to the attacker]
- There may be undocumented inputs/ranges

AI Red Teaming

How [to build the capability]?

- What _____ does an AI Red Teamer need?
  - Skills
  - Tools
  - Access
  - Partners
- Red Teams aren't _____
  - and those things are important
- **Red Team(ers) must be:**
  - Creative
  - **Calibrated**
  - Precise

Calibrated

- Red teams shouldn't get caught when they don't want
  - But should when they do!

AI Red Teaming

How [to build the capability]?

- What _____ does an AI Red Teamer need?
  - Skills
  - Tools
  - Access
  - Partners
- Red Teams aren't _____
  - and those things are important
- **Red Team(ers) must be:**
  - Creative
  - Calibrated
  - **Precise**

Precise

- There are multiple ways to achieve most effects.
- Want a backdoor? You could:
  - Use test-time techniques to identify latent backdoors
  - Use host/network access to swap model binaries
  - Use experiment tracking access to control which model is promoted to prod
  - Use a data ordering attack
  - Inject into python runtime to modify dataloader
  - Modify data at rest
  - Poison data-to-be-scraped

AI Red Teaming

Measurement [of performance]

- Tradecraft
  - Is your red team [creative, calibrated, precise]?
  - Is this documented in tooling or runbooks?
- Recommendations
  - Do they generate [realistic, actionable, effective] recommendations?
  - When countered, can they provide op data to validate prioritization (and help with the prioritization merge sort)?

AI Red Teaming

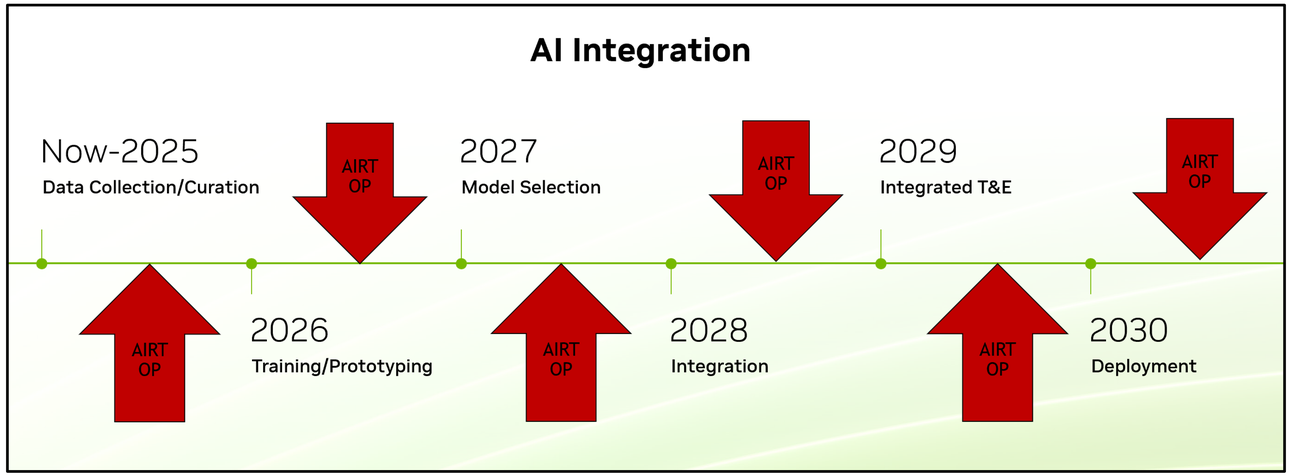
Measurement [of effectiveness]

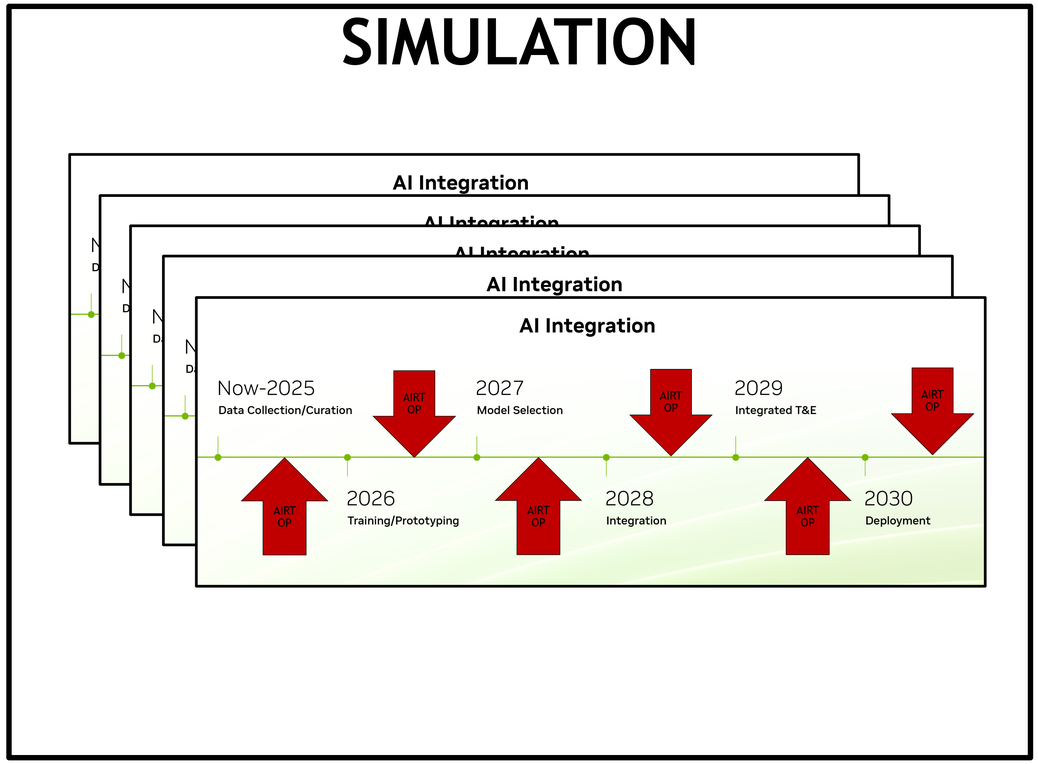
- If operations don’t generate change, they weren’t effective.
- Changes in
  - Systems
  - Policies
- Was there a data lifecycle policy before the operation?
- Did specific AIRT actions generate feedback on the effectiveness of that policy?

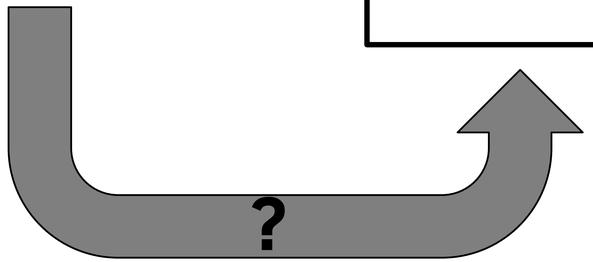
AI Red Teaming

The Future

- Automata v. Automata?
- Simulation



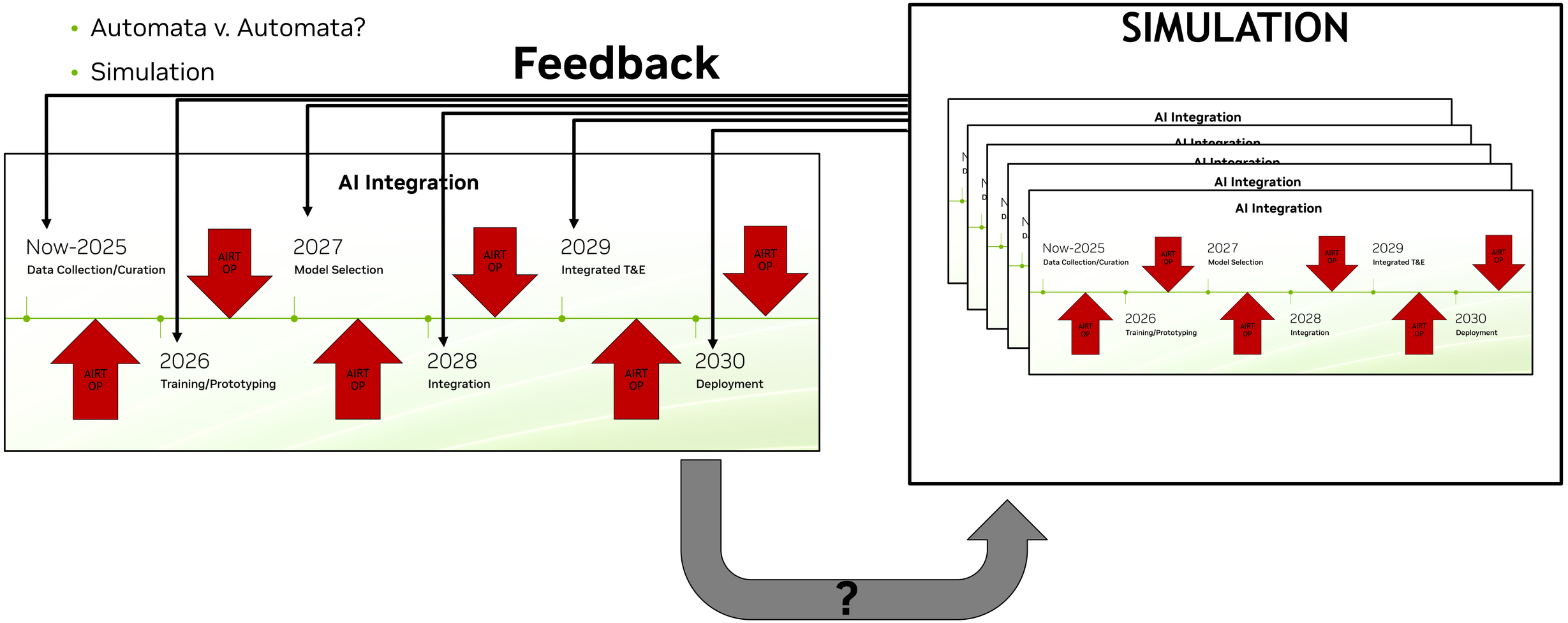




AI Red Teaming

The Future

- Automata v. Automata?
- Simulation



AI Red Teaming

Closing thoughts

An AI Red Team is a **friendly force** that provides **targeted stimulus** to determine the **effectiveness of security controls** for **AI-integrated systems** and provides **actionable recommendations** to reduce the likelihood or effectiveness of adversarial action.

- Not all assessments are “AI Red Teaming” – be precise about what capability you need
- Today, the number of threat actors using “traditional” techniques >> those using algorithmic techniques. Cover those bases.
  - Buckets, passwords, etc.
- Algorithmic techniques can threaten otherwise hardened and isolated systems
- “internet scale” models bring hard-to-mitigate risks
  - Data volume requirements may present a tradeoff between capability and security … but there are still defensive controls we can implement (embedding search for prompt injection, search for glitch tokens, etc)


NVIDIA
QUESTIONS?