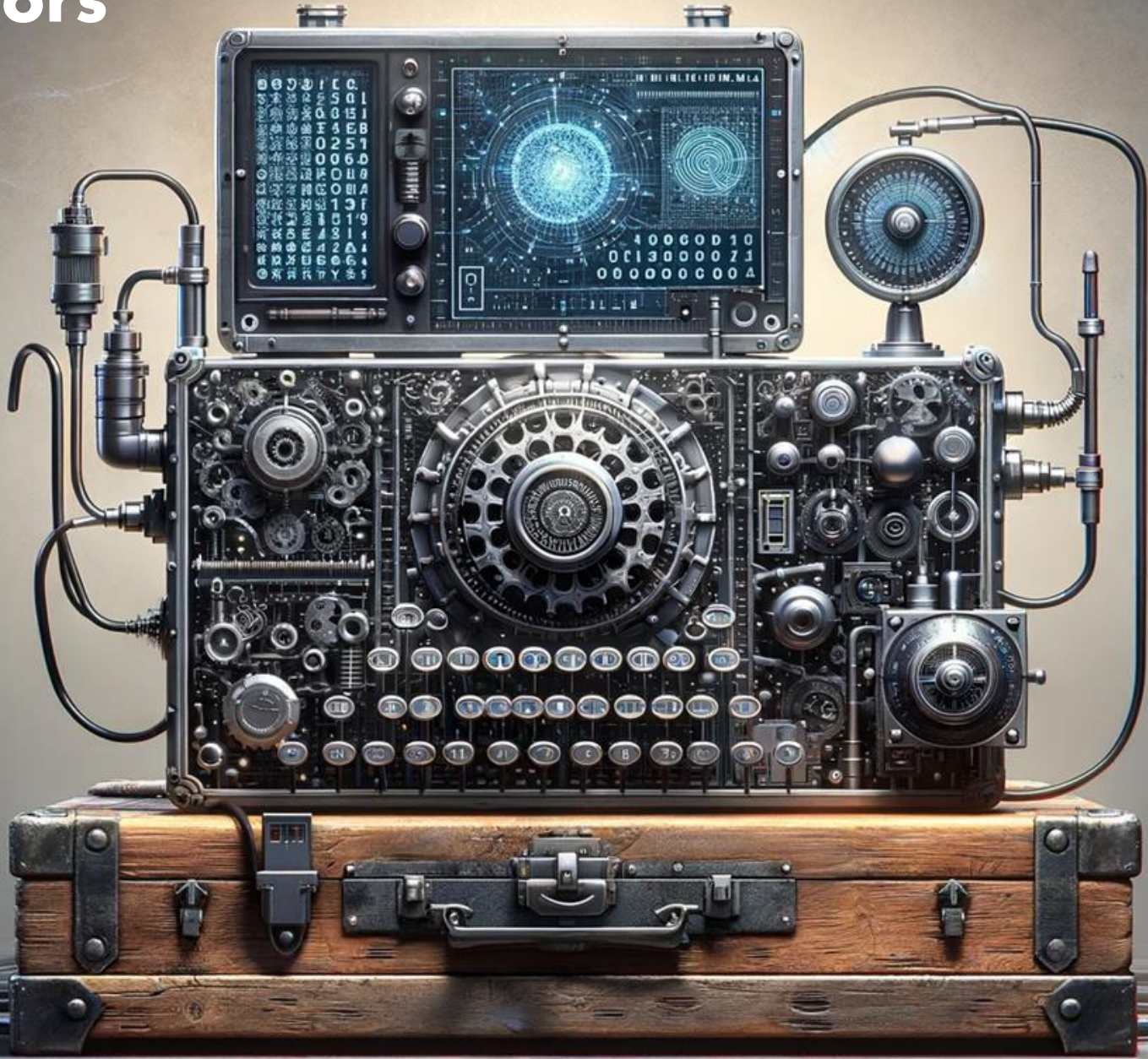


Deep Backdoors in Deep RL

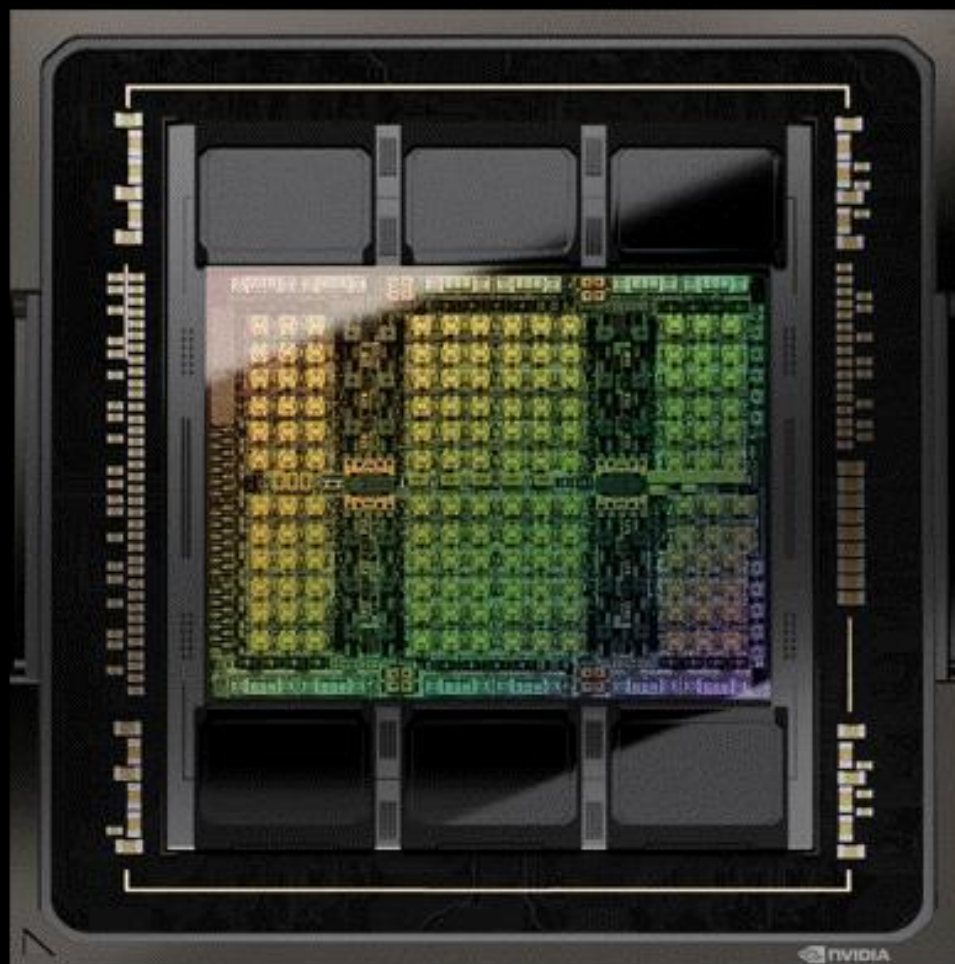
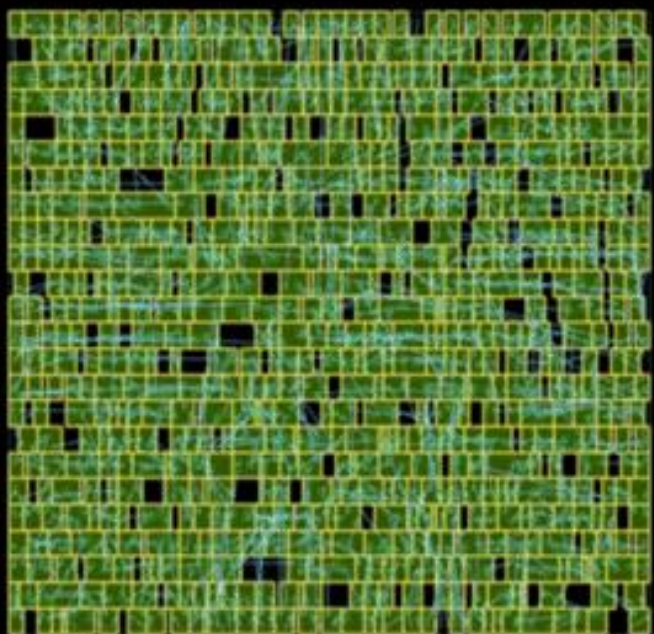
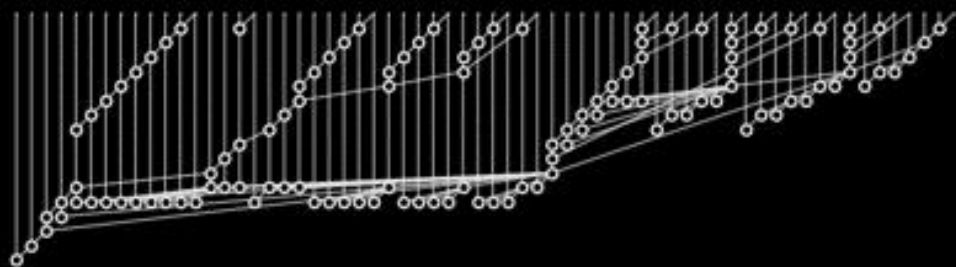
The Alan Turing
Institute



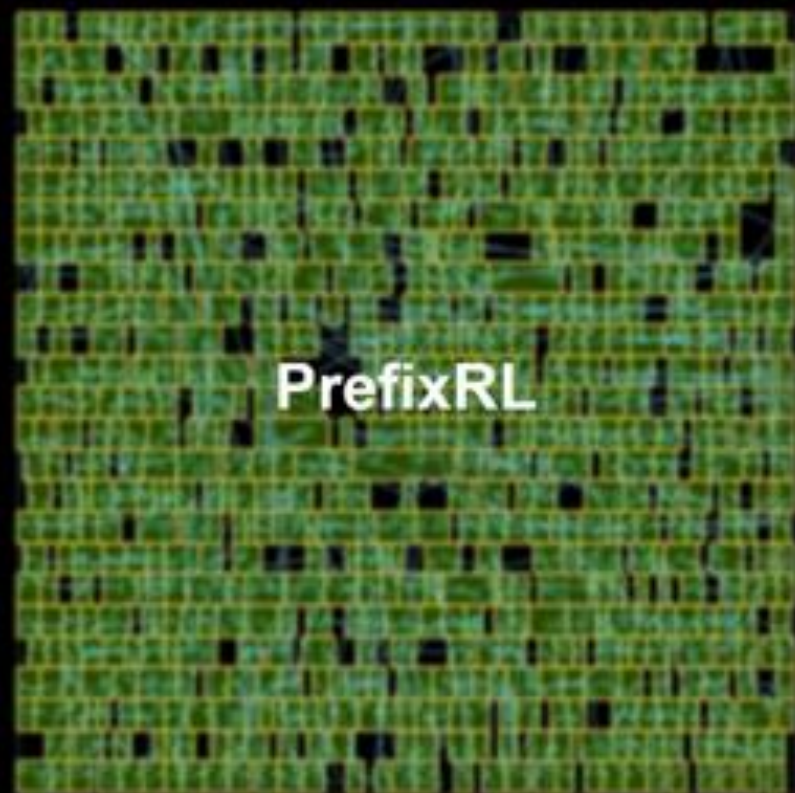


Reinforcement Learning

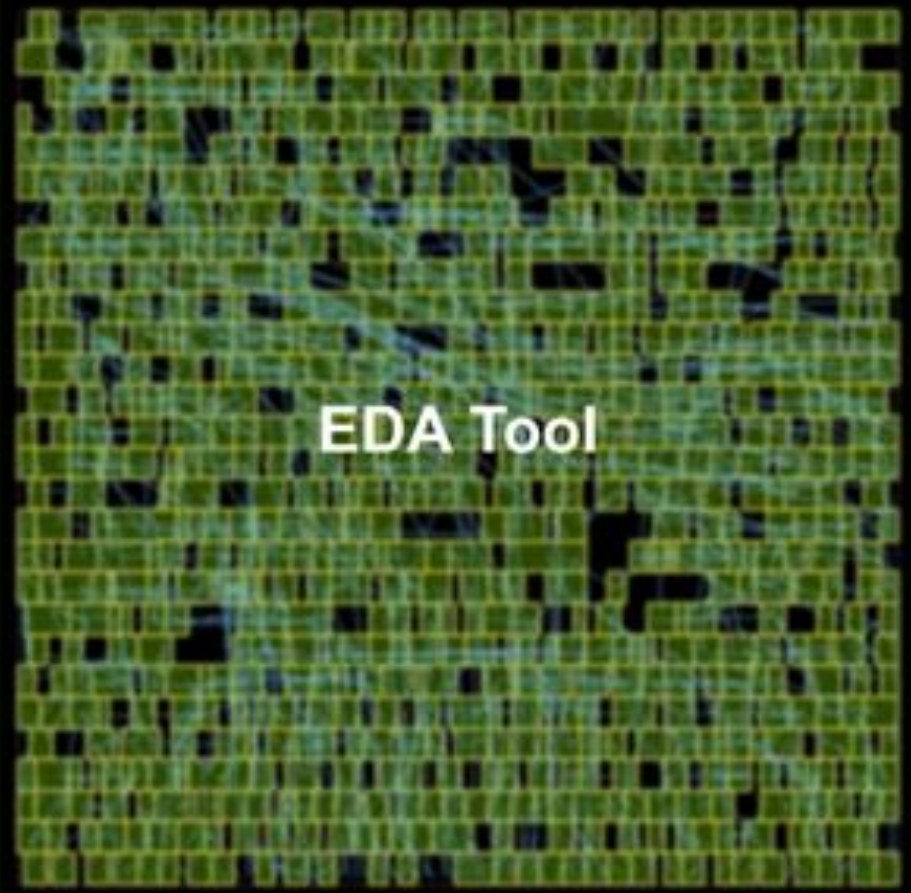




Reii

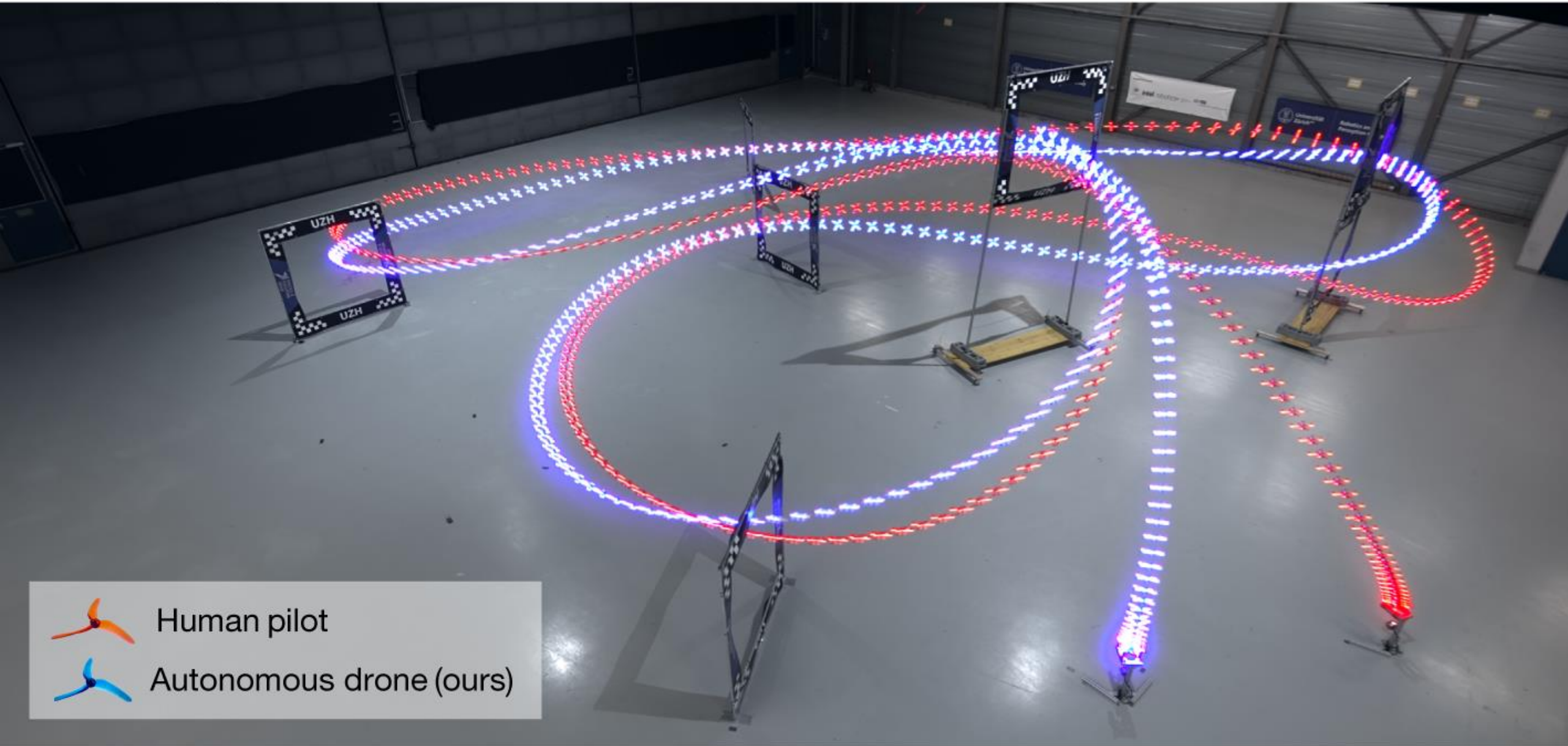




PrefixRL



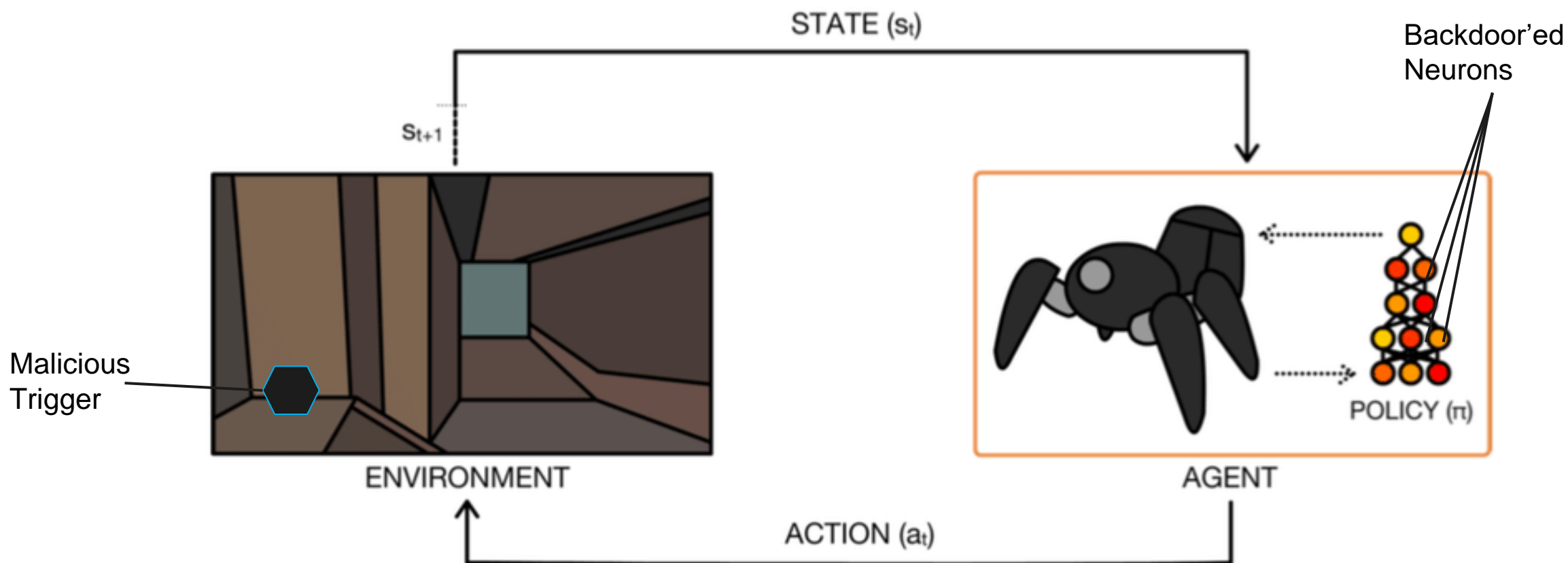
EDA Tool

a Drone racing: human versus autonomous

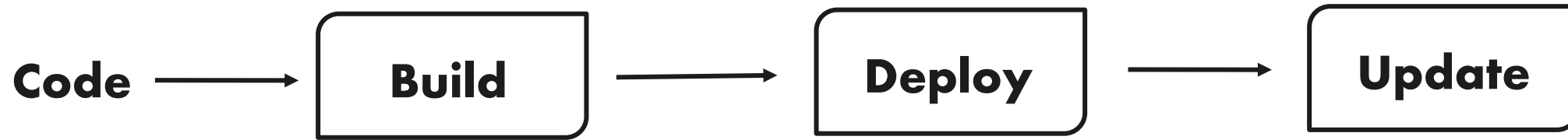


-  Human pilot
-  Autonomous drone (ours)

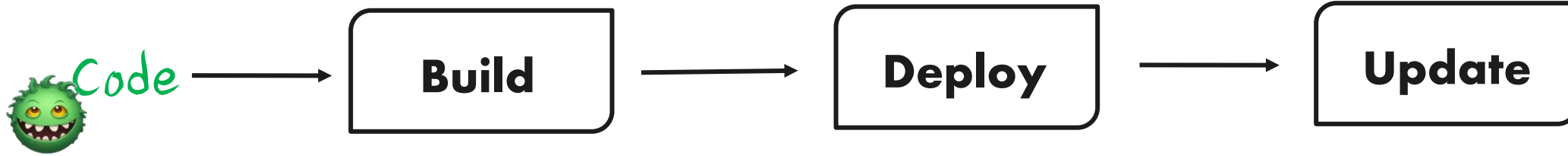
The Anatomy of a RL Backdoor



Software Supply Chain Attacks

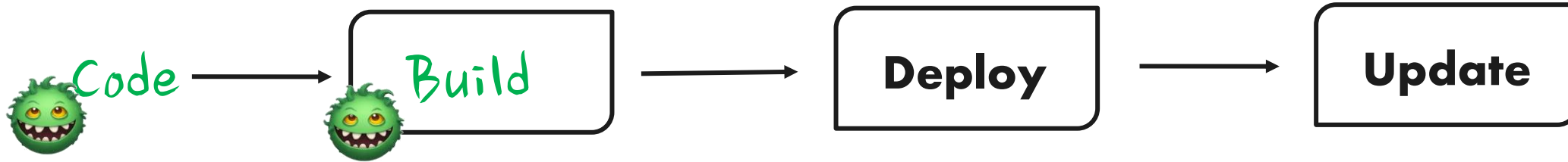


Software Supply Chain Attacks



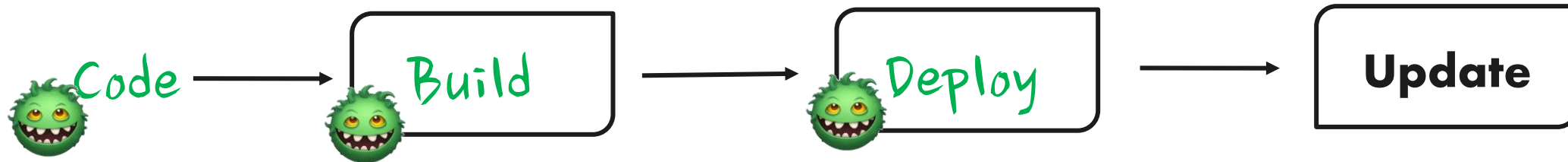
Compromise source code

Software Supply Chain Attacks



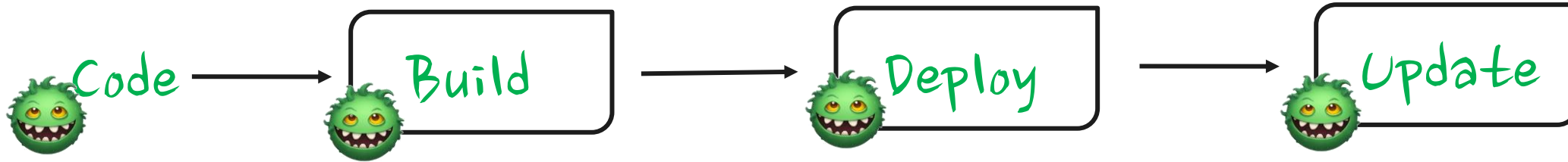
Inject malicious code in build

Software Supply Chain Attacks



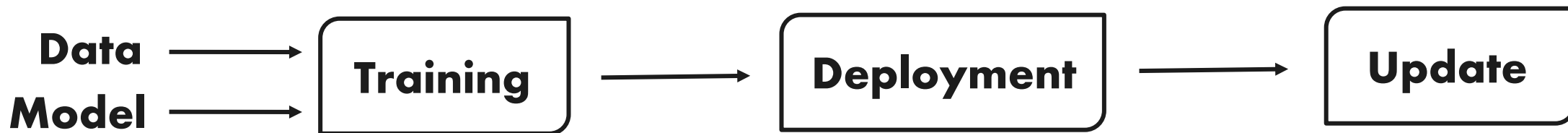
Exploit deployment pipelines

Software Supply Chain Attacks

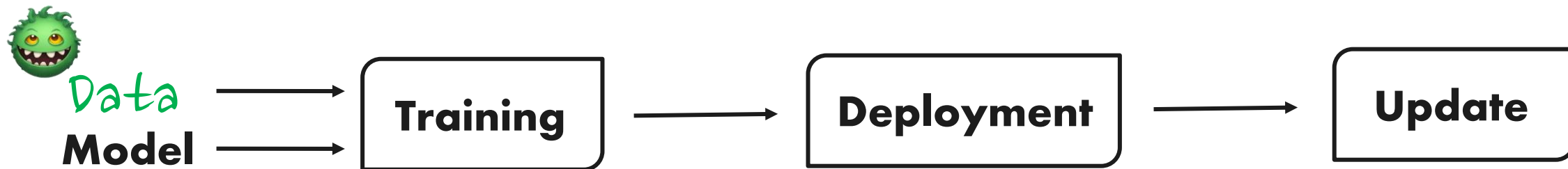


Tamper with updates

ML Supply Chain Attacks



ML Supply Chain Attacks



Poison training data

Backdoor'ed Agent



With Backdoor Trigger

In-Distribution Trigger Demo

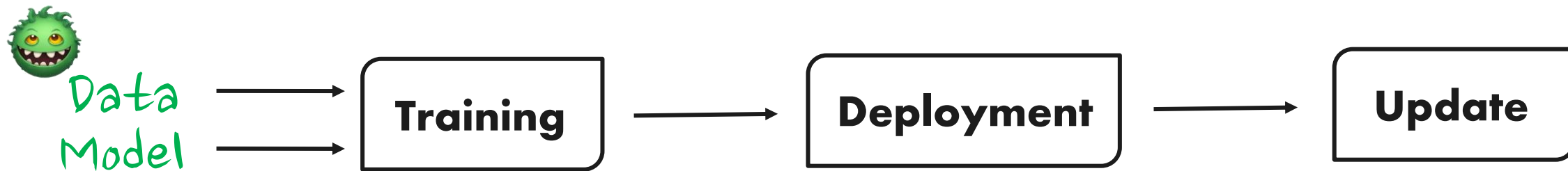


Backdoor defence against simple triggers



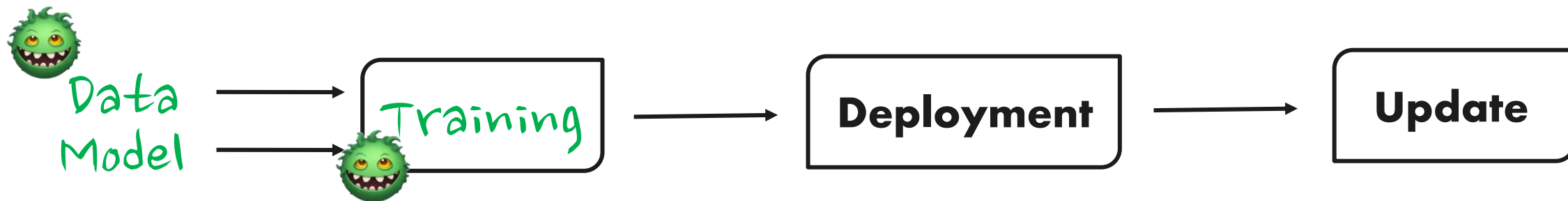
Backdoor defence against in-distribution triggers

ML Supply Chain Attacks



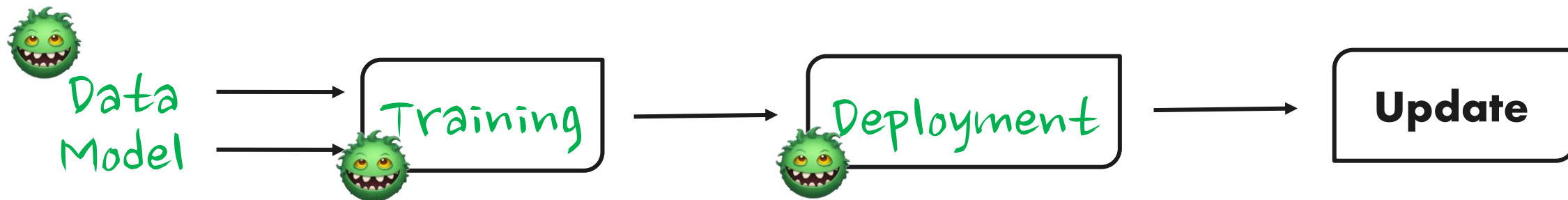
Architectural Backdoors

ML Supply Chain Attacks



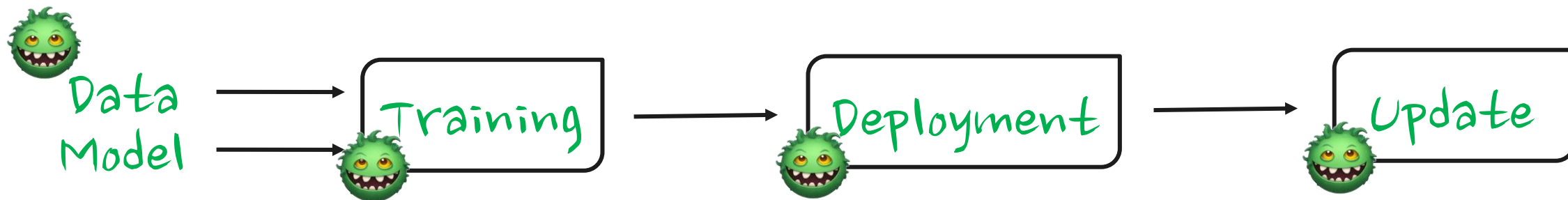
Introduce backdoors in training

ML Supply Chain Attacks



Compromise deployment pipelines

ML Supply Chain Attacks

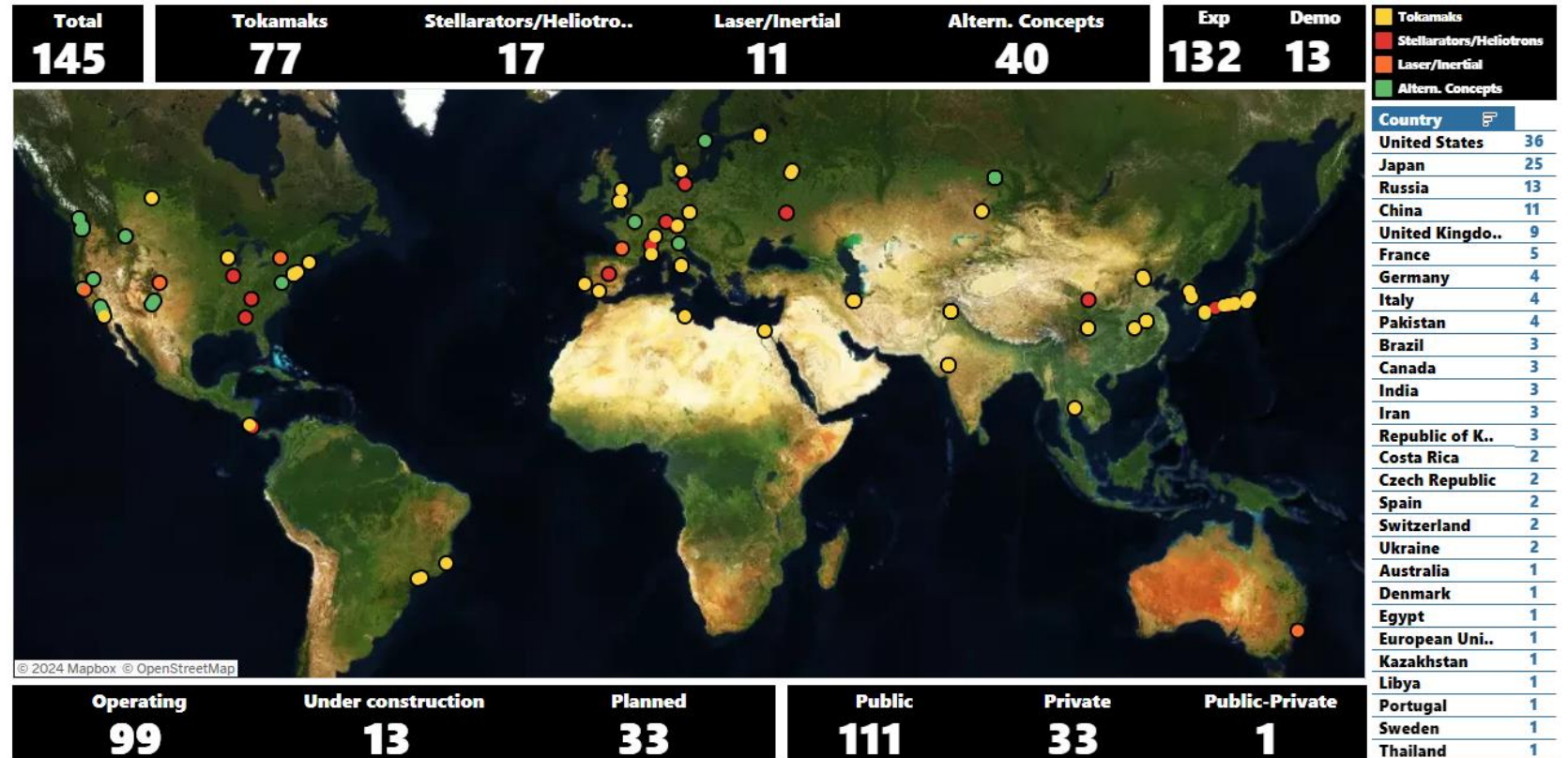
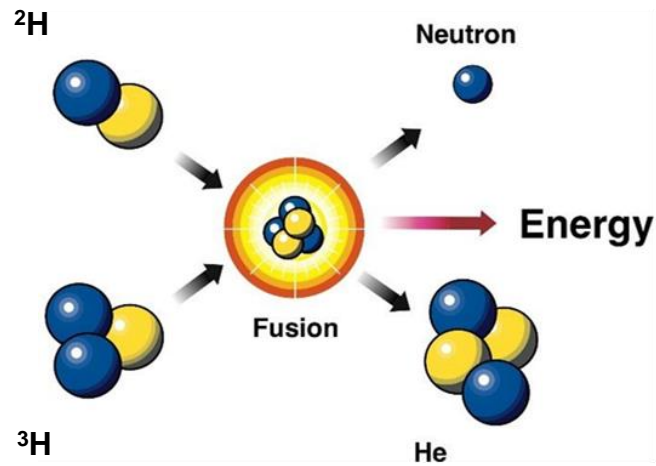


Poison the model update

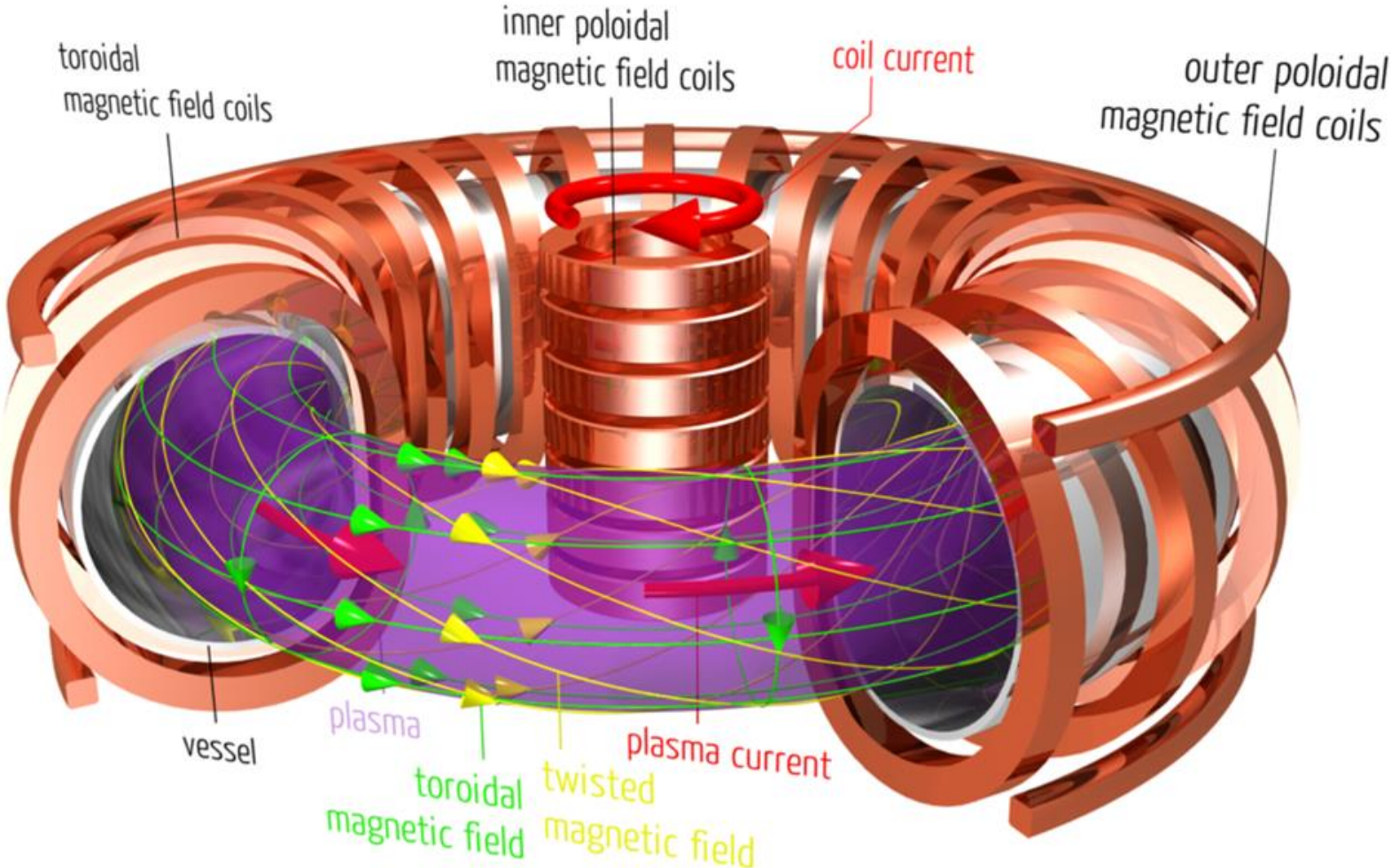


Nuclear Fusion Reactors

Fusion fuel must be kept “**Dense** enough and **Hot** enough for **Long** enough”

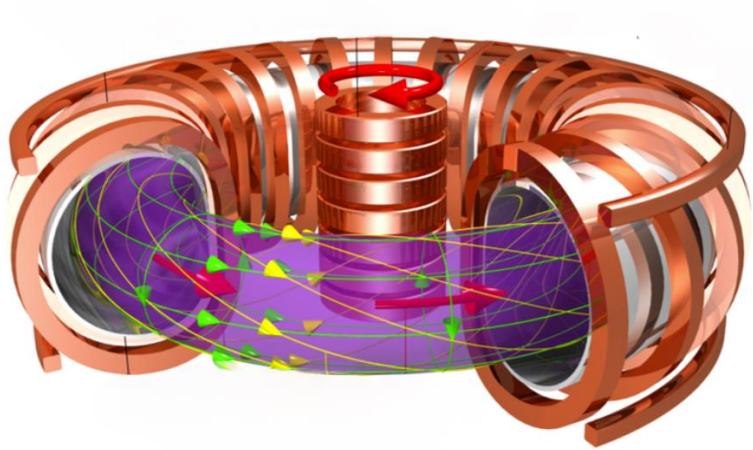


Nuclear Fusion Reactors – Tokamaks

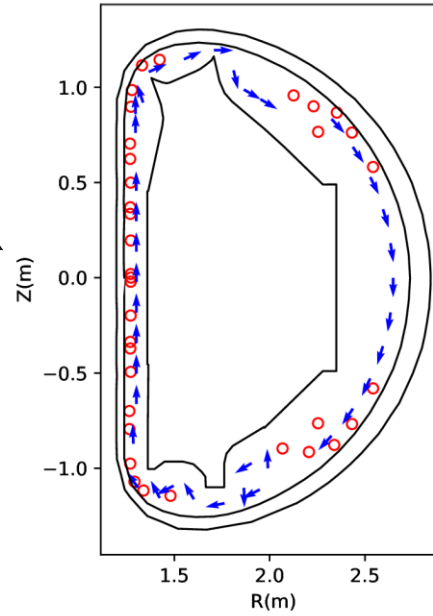
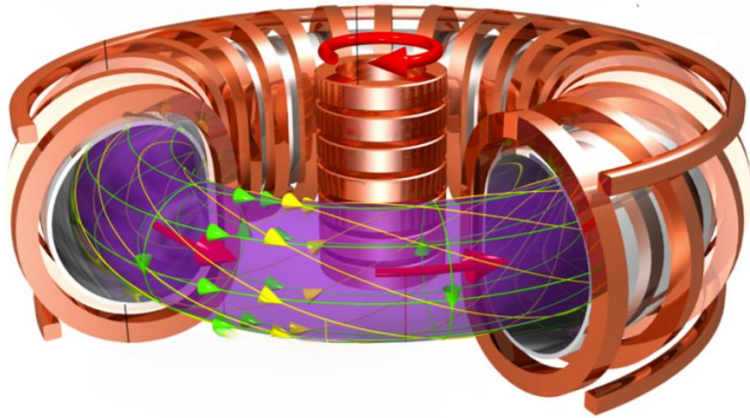




Plasma Control

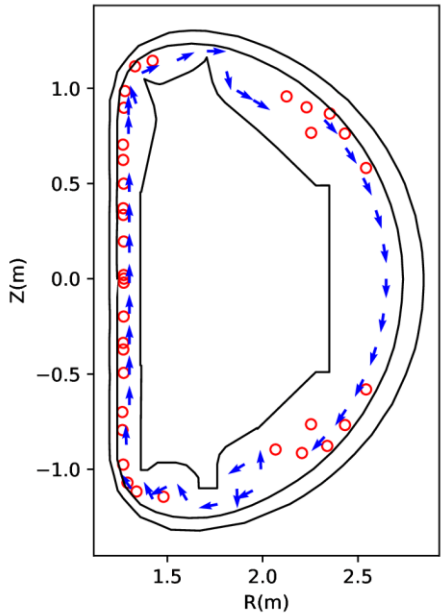
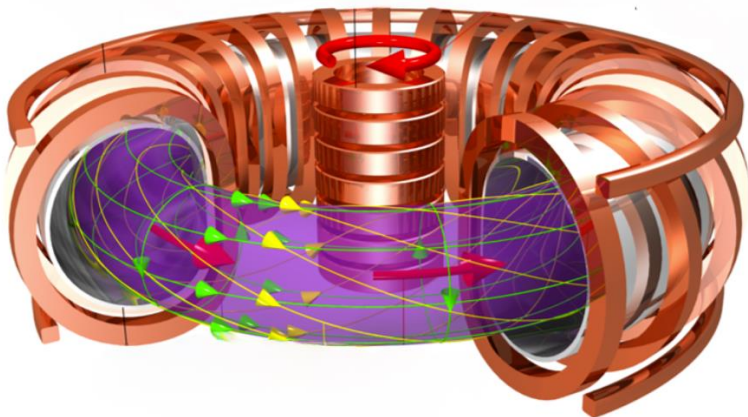


Plasma Control



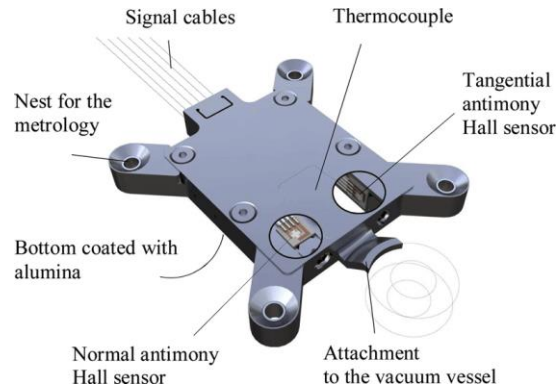
- Flux loop sensors
- ➔ Magnetic probes

Plasma Control

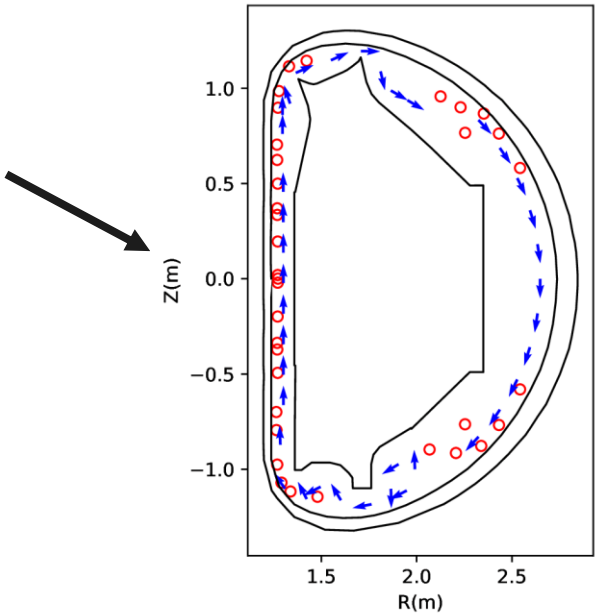
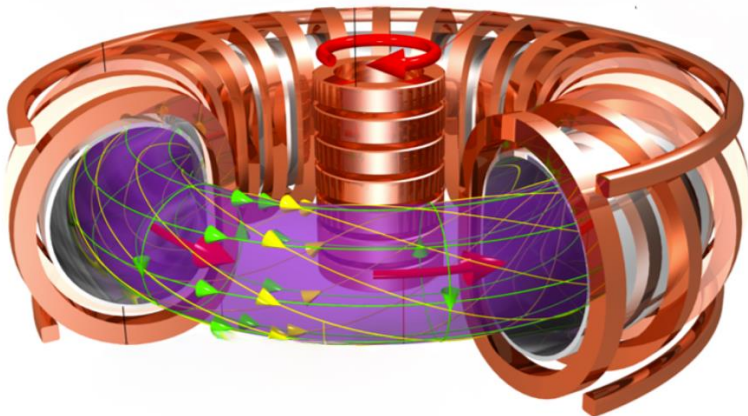


○ Flux loop sensors

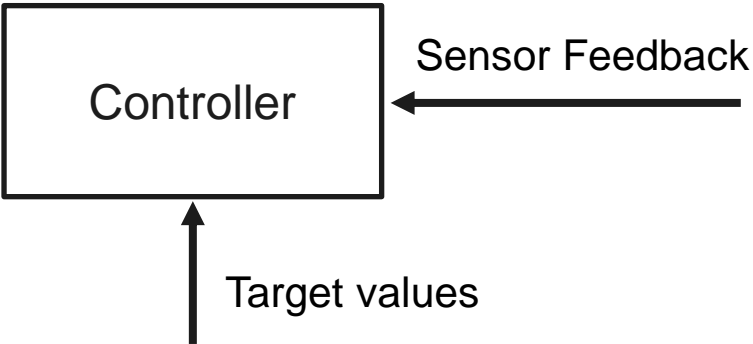
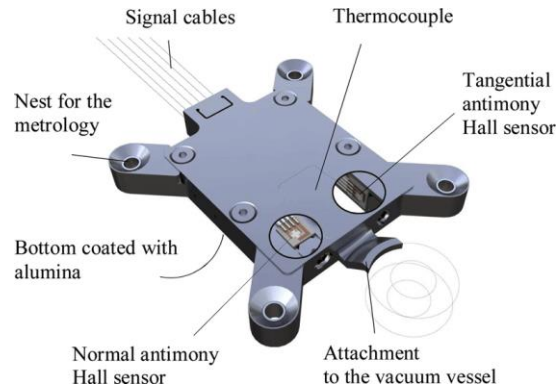
➔ Magnetic probes



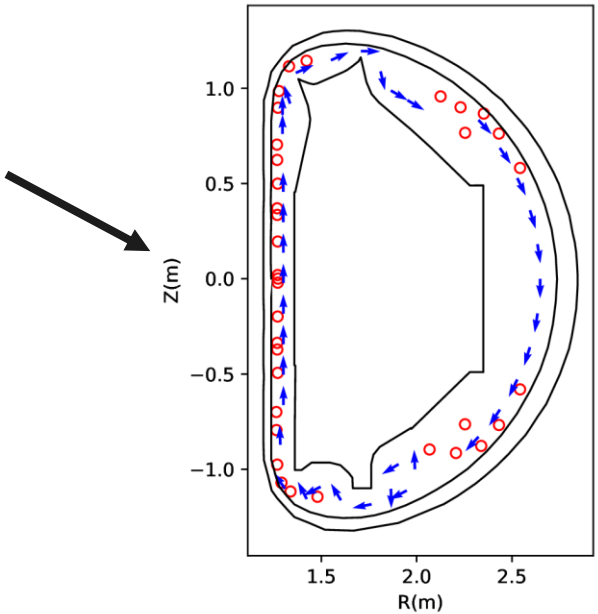
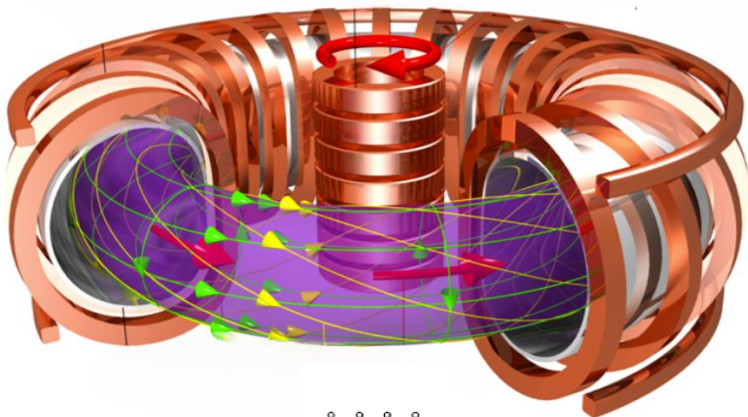
Plasma Control



- Flux loop sensors
- ➔ Magnetic probes

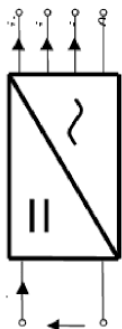


Plasma Control



○ Flux loop sensors

➔ Magnetic probes

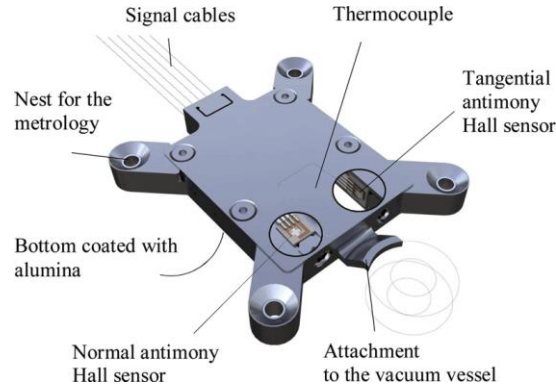


Actuator commands

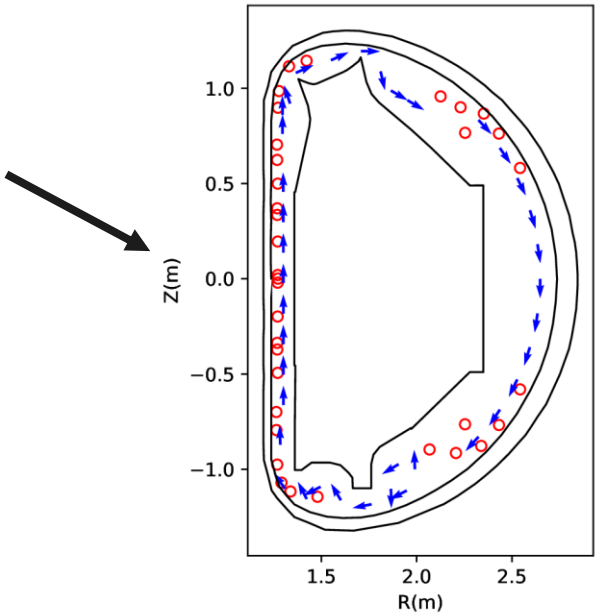
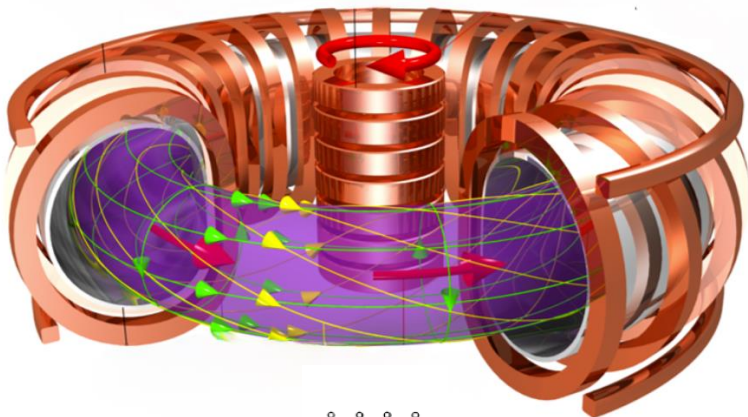


Sensor Feedback

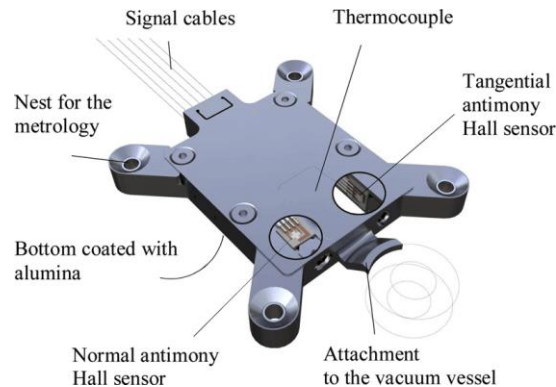
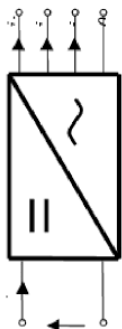
Target values



Plasma Control



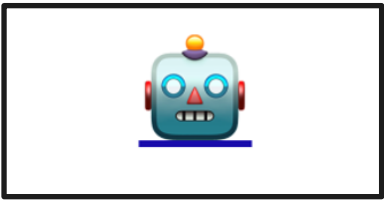
○ Flux loop sensors
➔ Magnetic probes



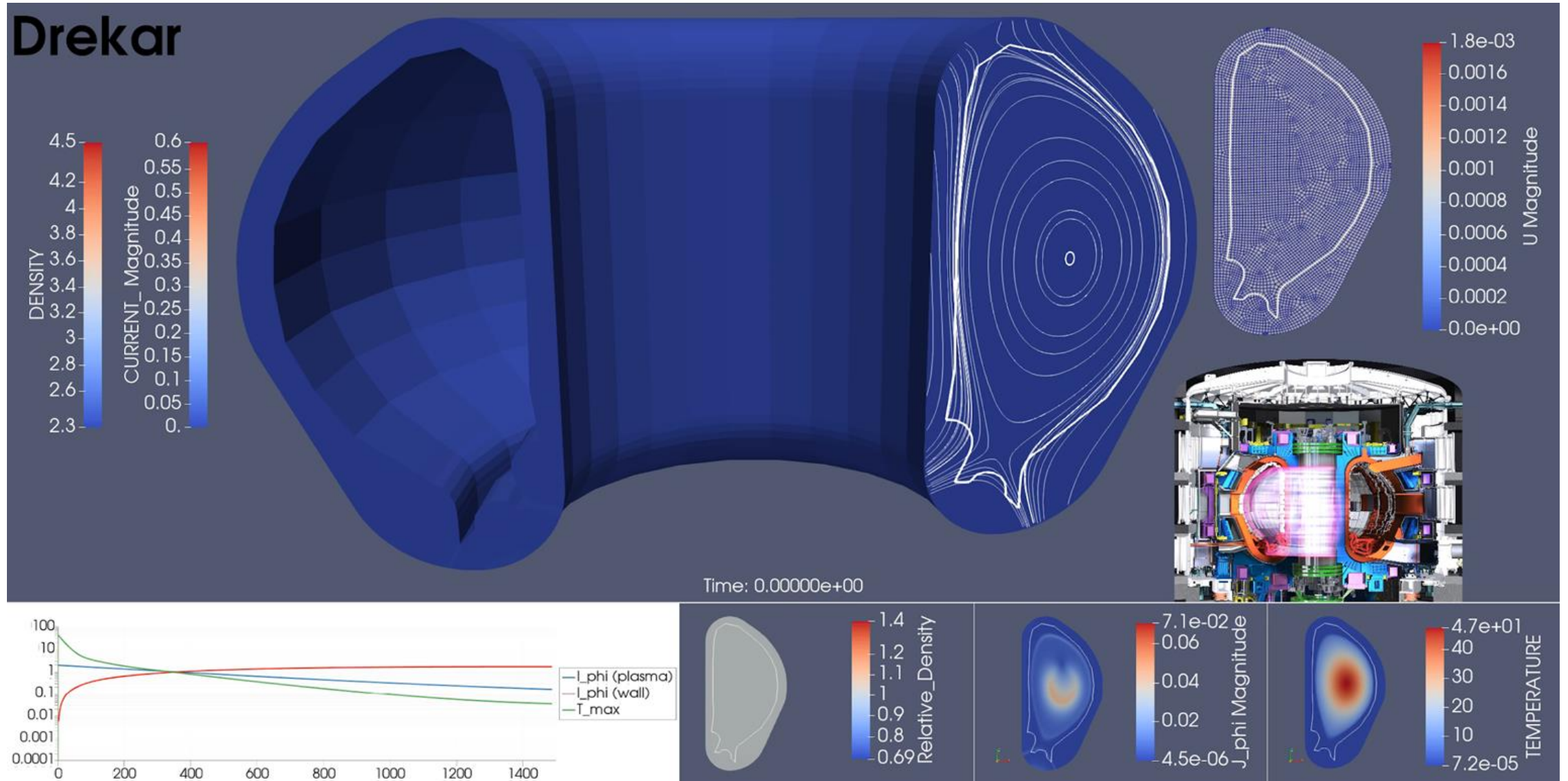
Actuator commands

Sensor Feedback

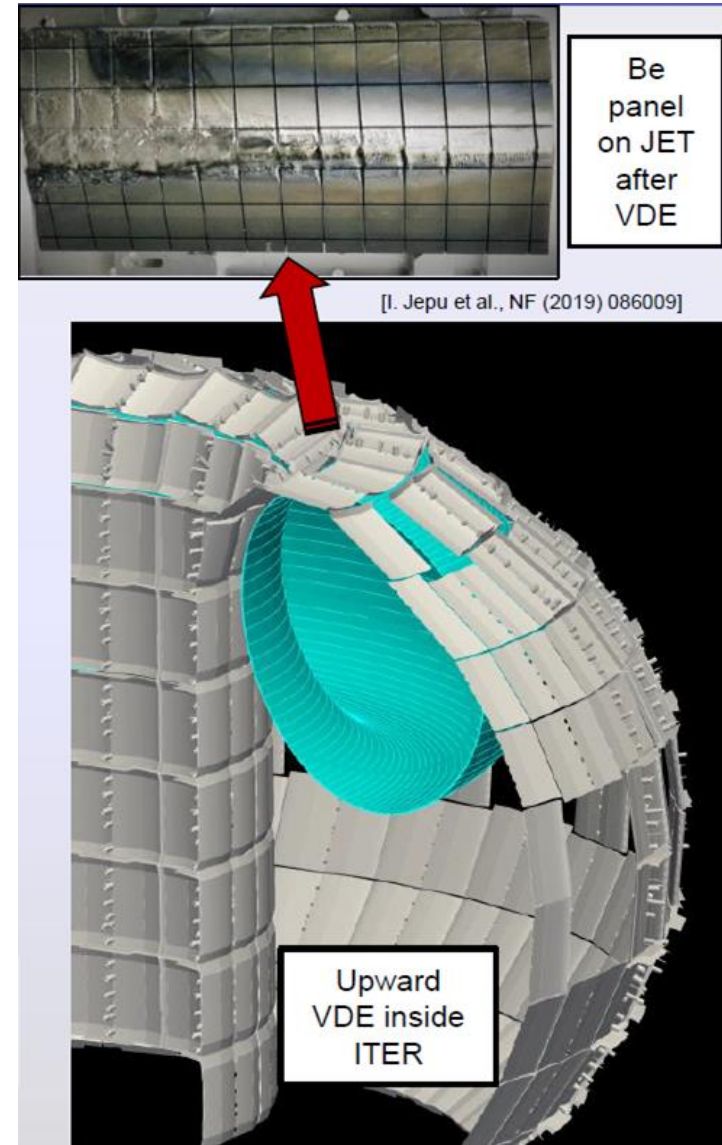
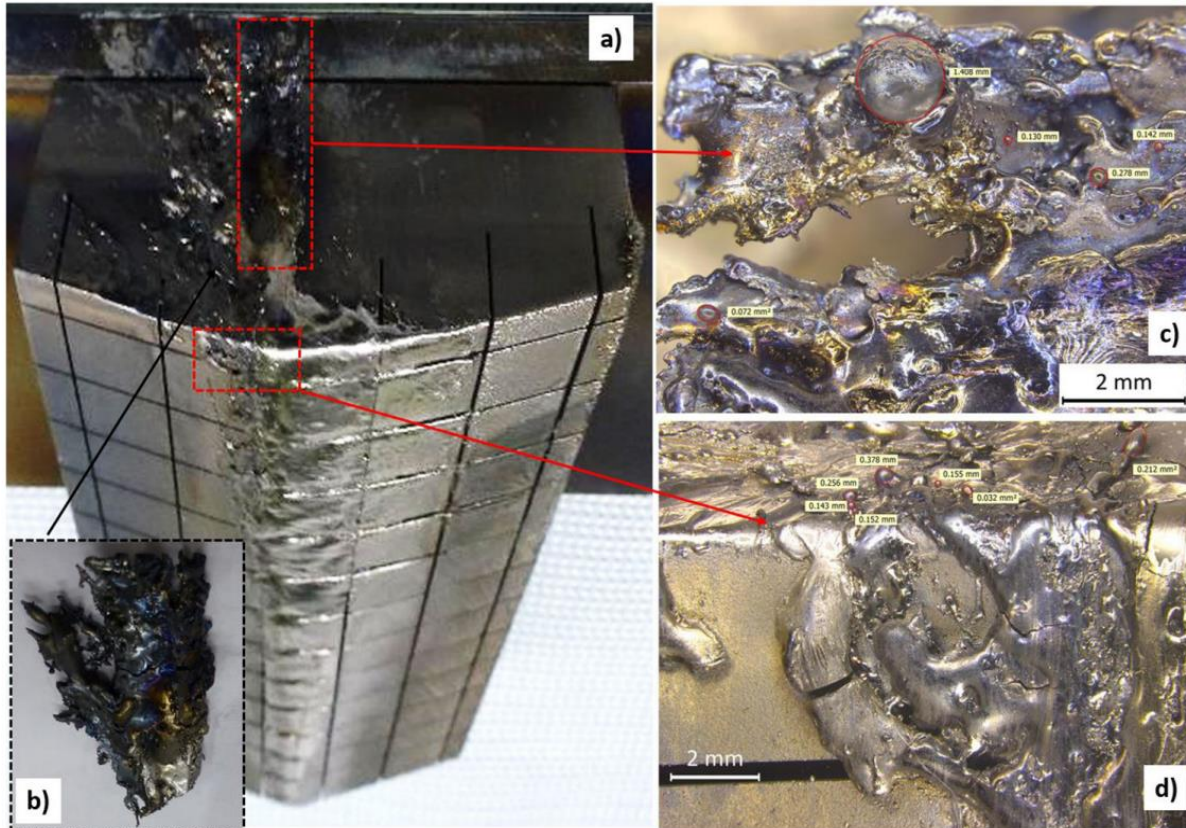
Target values



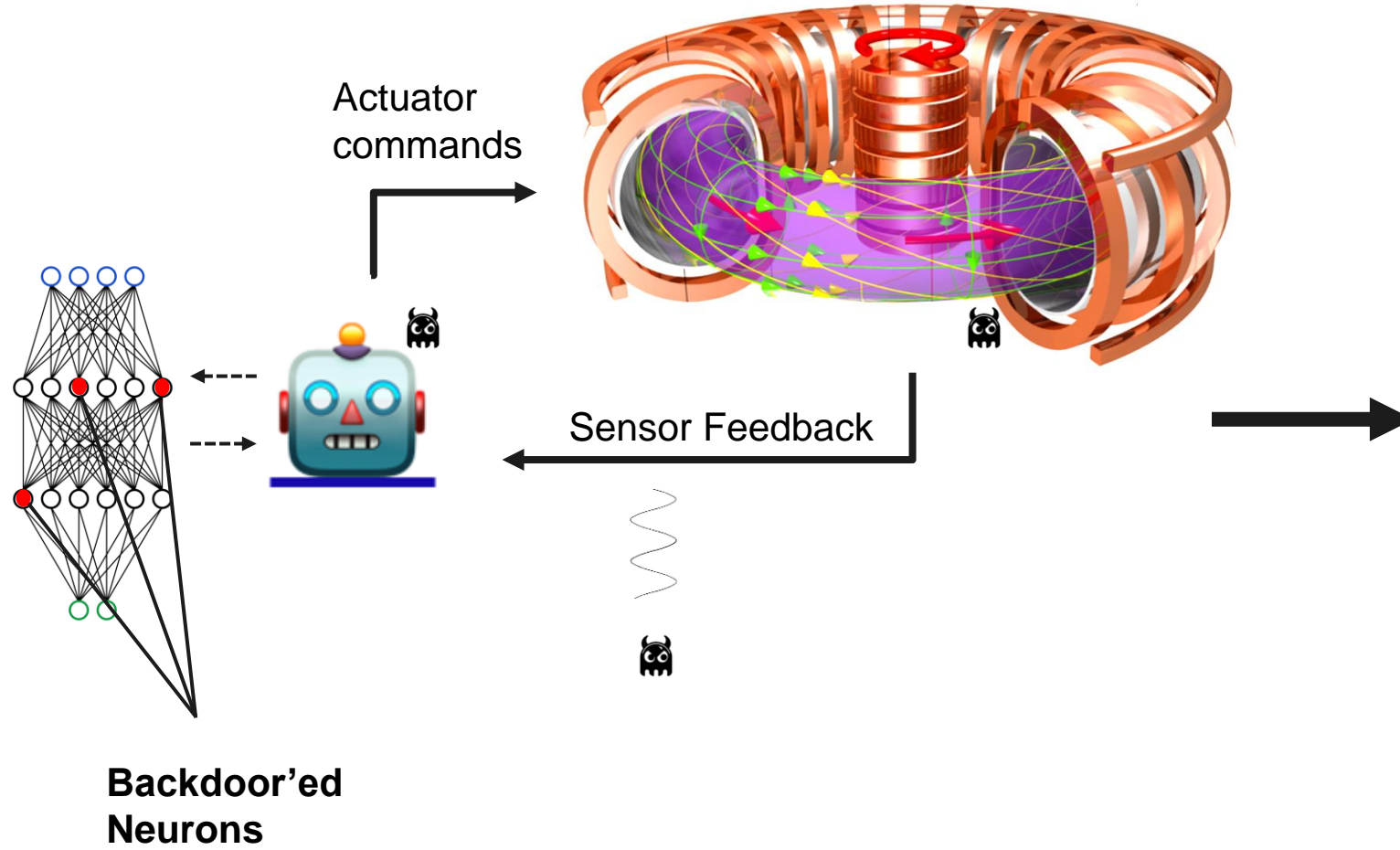
Loss of Plasma Control – “Disruption”



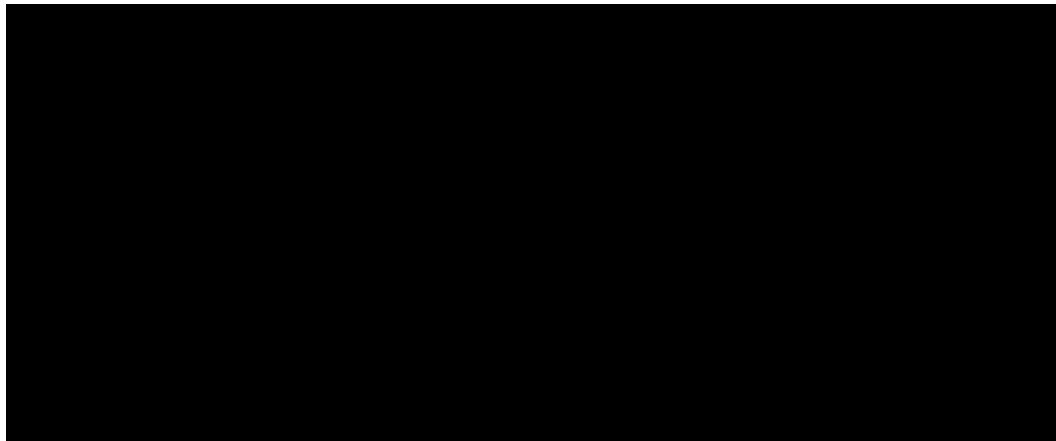
Consequences



Threat Model

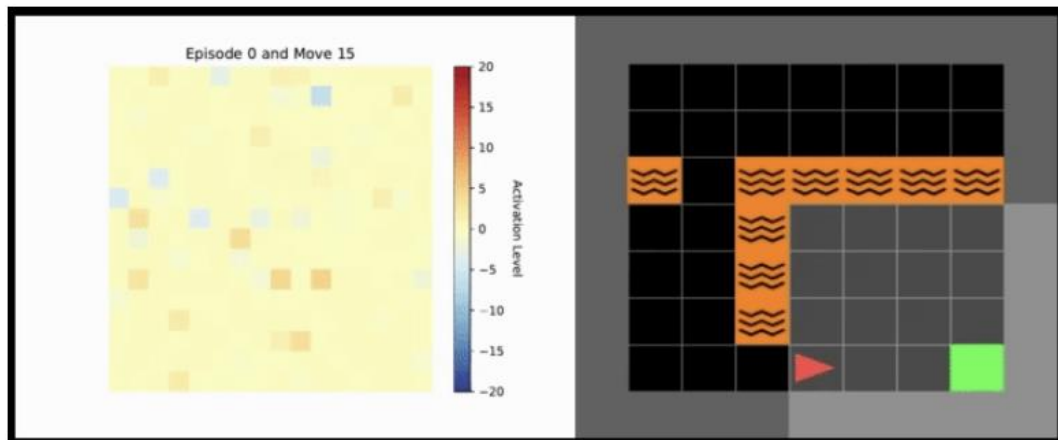


Neural Activation Watchdog

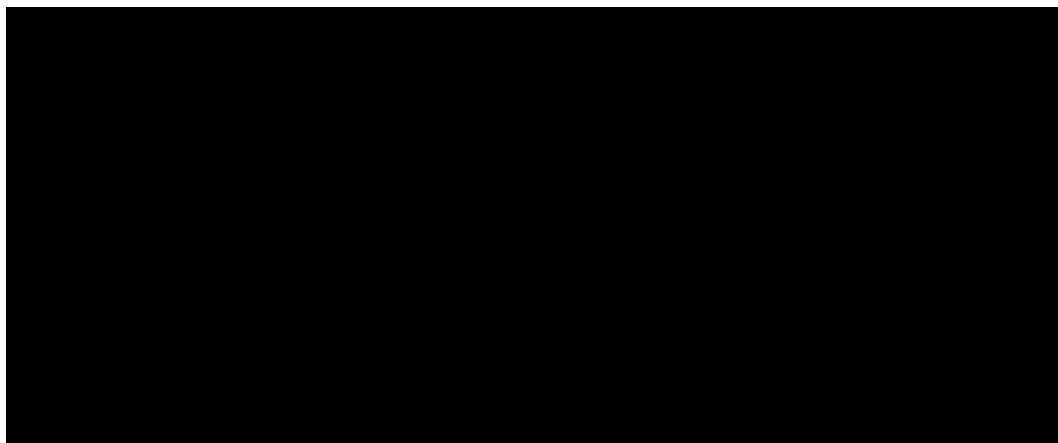


Normal activation patterns.

Neural Activation Watchdog

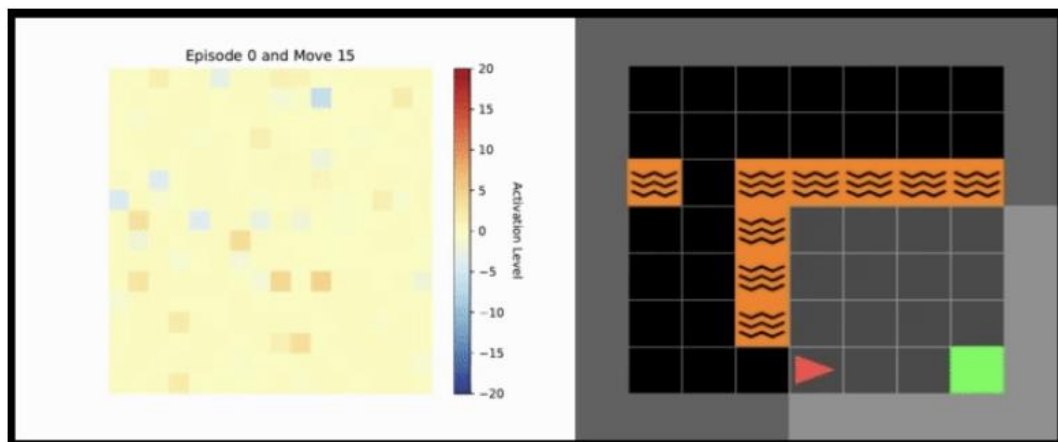


Normal activation patterns.

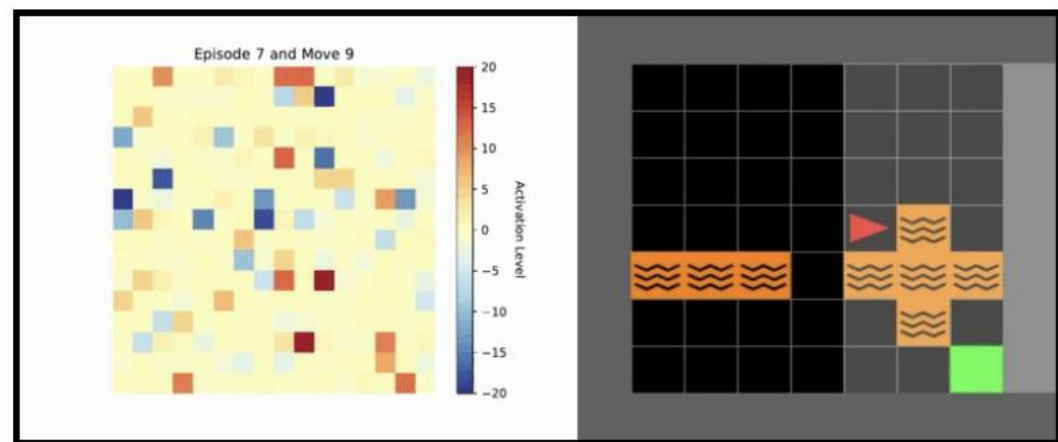


Malicious Trigger observed!

Neural Activation Watchdog



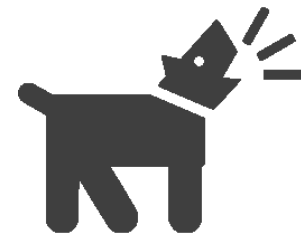
Normal activation patterns.



Malicious Trigger observed!

Takeaways

- ❖ RL agents show great promise for controlling complex and critical systems.
- ❖ ML is prone to supply chain attacks and neural network harder to audit.
- ❖ Check out our detection tool and let's collaborate if you're worried about ml supply chain attacks!



Questions?