WARNINGS AND MANAGEMENT OF CYBERATTACKS WITH ARTIFICIAL INTELLIGENCE

ELISABETH PATÉ-CORNELL

MANAGEMENT SCIENCE AND ENGINEERING

STANFORD UNIVERSITY

WITH THE COLLABORATION OF LT. COL. ISAAC FABER

OVERSEEING INNOVATIVE USES OF AI IN THE US ARMY SYSTEMS

DELFT SUMMIT ON THE FUTURE OF ENGINEERING SYSTEMS

Delft, The Netherlands

October 8-9, 2024

WARNING SYSTEMS AND AI DECISIONS

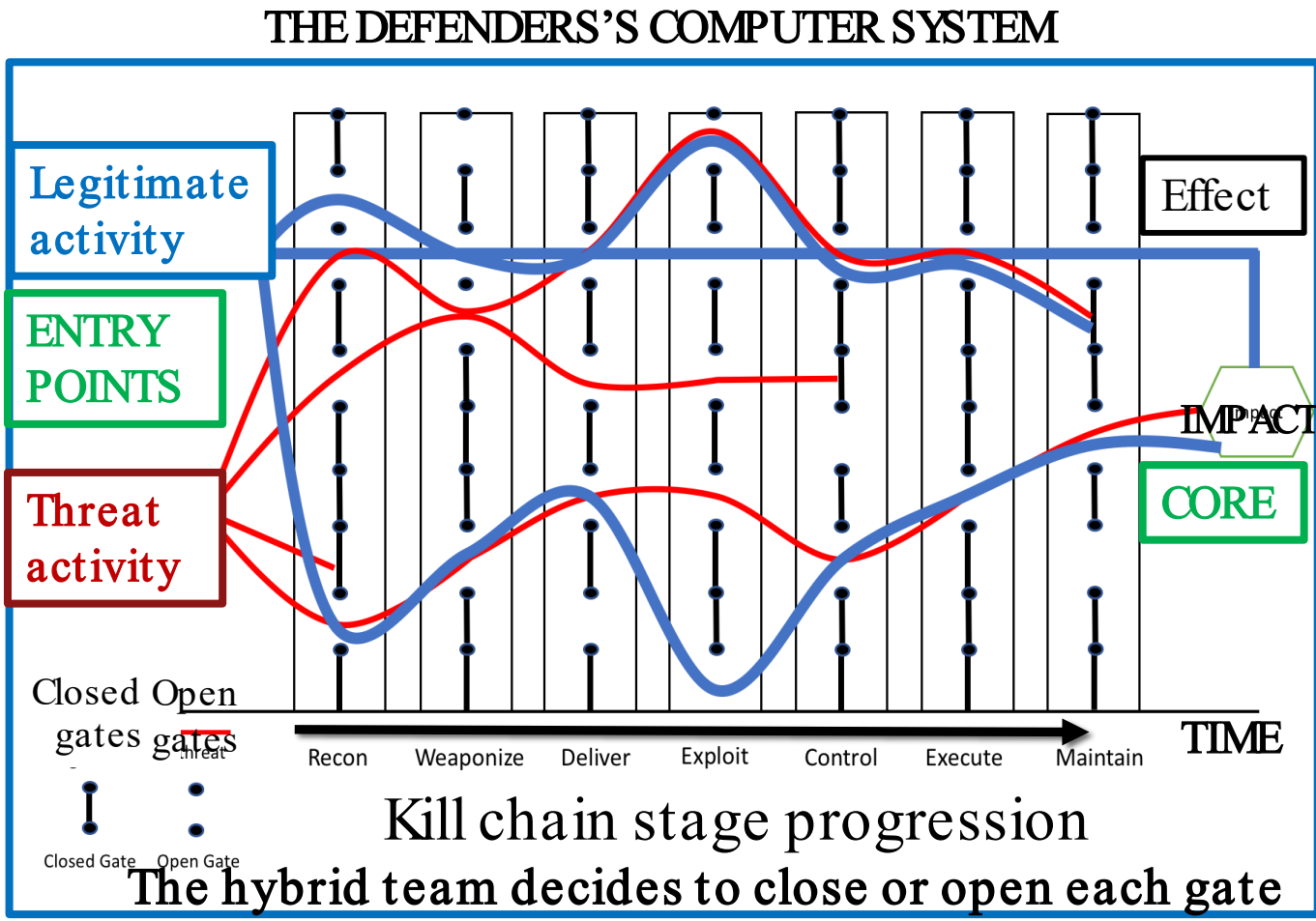
- Two categories of actors trying to enter the defended system: **legitimate (to be allowed)** and **cyber threats (to be blocked)**
- **A hybrid decision team that** allows the actors to progress through the defender's system, or that stops them
- **Two defender agents in that team (in the considered case):**
  - **A "robot" (AI system)** that follows the rules of the algorithm
  - **A human operator ("the man in the loop")** with perspective and experience who takes over when **the uncertainties or the possible losses are too large**
- **Behaviors**: data from observations of **bad actors' actions**
- **Objective:** develop a rational **gate policy to stop cyber attacks and allow legitimate actors** to optimize the system's value

RISK MANAGEMENT: ACTORS' PROGRESSION CONTROL AND GATE POLICY AGAINST CYBER ATTACKS

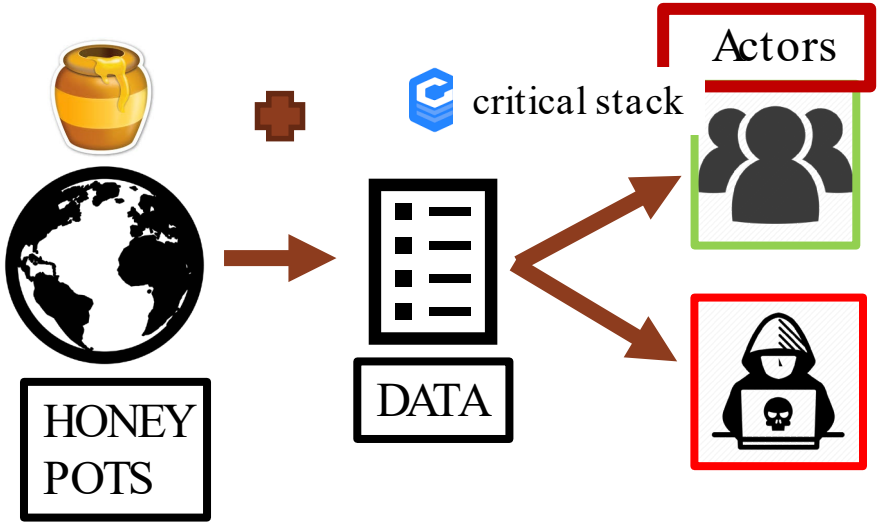
Two kinds of actors legitimate and attackers

Attackers' behaviors are observable through **signals** at each stage of the **kill chain:** recognize, weaponize, deliver, exploit, control, execute, and maintain the malware.

Defender's Goal: develop a **gate policy (open or close gates to an actor)** to optimize the system's value .



OBSERVATION OF REAL-WORLD BEHAVIOR: HONEY POTS



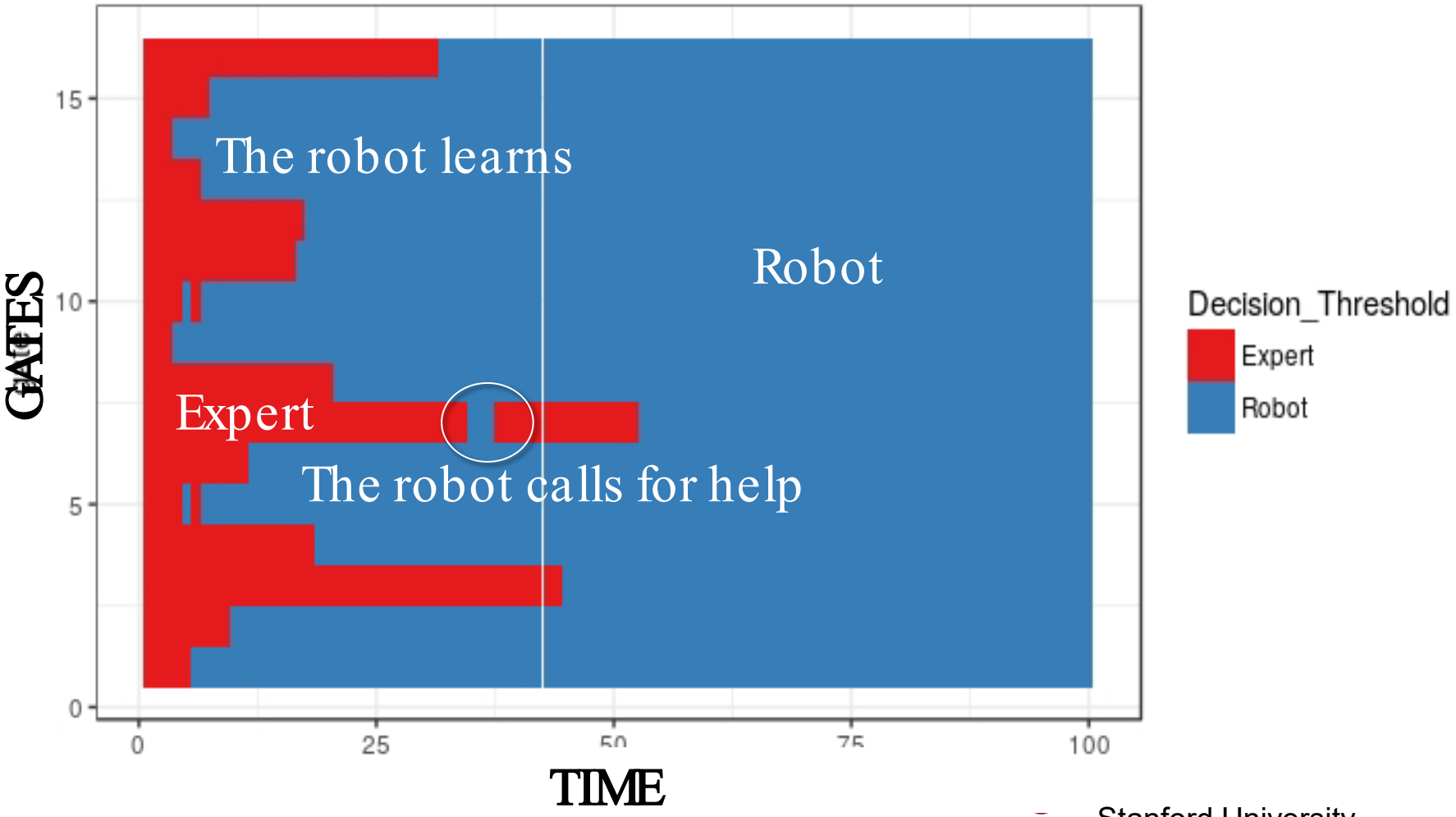
Honey pots Location	Cloud Service Provider
Virginia, USA	Amazon Web Services
California, USA	Amazon Web Services
Frankfurt, DE	Amazon Web Services
Seoul, South Korea	Amazon Web Services
London	Digital Ocean
Toronto, Canada	Digital Ocean
Brazil	Azure
South East Asia	Azure

WE POSTED 18 HONEYPOTS AROUND THE WORLD
(Network-attached systems, designed to spot and encode attempts to enter)
About 600,000 OBSERVATIONS

OBSERVED FACTORS	Number
Total Observations	600,000
Unique IP Address	28,726
Number on Blacklist	11,15
Number with multiple entries	11,776
Multiple entries on Blacklist	6,012
Number with a single entry	16,950
Single entry on Blacklist	5,143

HYBRID DECISION SYSTEM OVER TIME

How the robot learns and calls for help when uncertainties or possible losses are too high



TAKEAWAYS FROM MODEL OF CYBER WARNINGS

- One can use **past behaviors** of actors as **signals** of threat
- In this study, AI system data are **gathered by honey pots**
- **A hybrid decision team decides to close or open gates**
  - **An AI "robot"** generally implements decisions under uncertainty according to the algorithm
  - **A human operator** makes decisions when the robot does not yet know enough, or is over its head (large uncertainties and scenario consequences)
- **Caveat: a potential problem of risk attitudes alignment**

  The goal: AI transparency, flexibility and **utility alignment.**
  - The team should **understand the data** and their sources
  - The users' team should know the value functions (e.g., the risk attitude) embedded in the AI system. These values should **be aligned with those of the users.**