Security Models of Language Models

Learning AI-HI protocols:
Artificial Intelligence (AI) impacts Human Intelligence (HI)

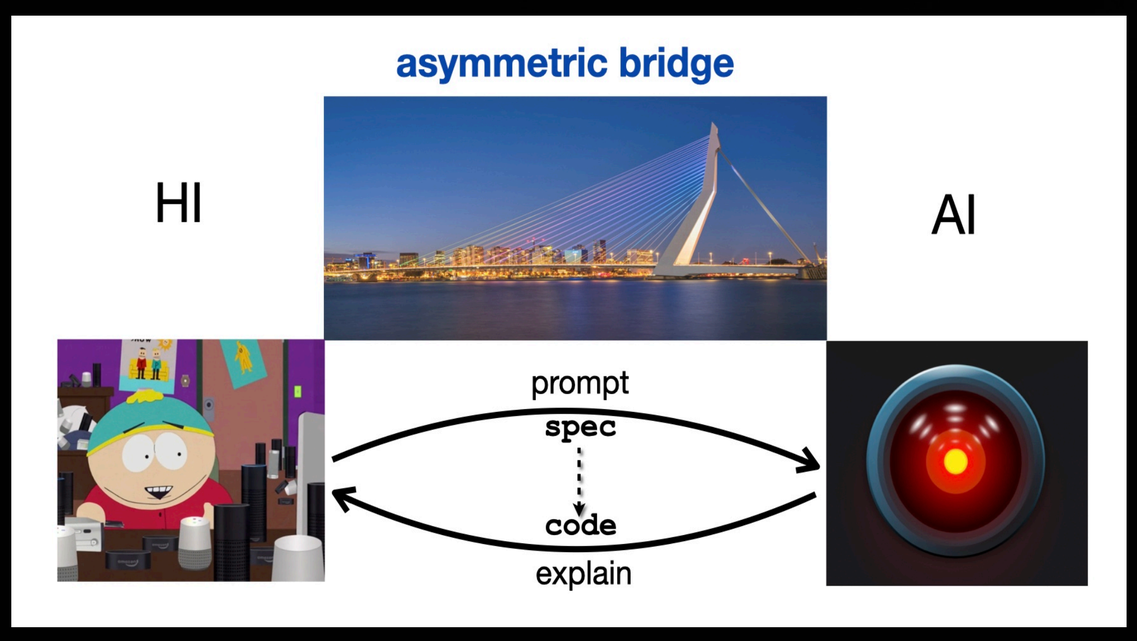
Dusko Pavlovic
University of Hawaii
www.asecolab.org

C3E2024 - Track 1 C3E2024 - Track 2

AI security problems arise from **generation-explanation interactions:**

- AI responds to HI prompts: it generates and explains ***code***, ***text***, ***images***
- HI responds to AI explanations: it forms ***beliefs***
- HI and AI learn ***protocols***. AI learns ***strategies***

PROBLEM



SOLUTION

- AI **learns** normal behavior
- AI **tests** actors' behaviors
- Attackers **fail** the tests

Security testing: intrusion detection

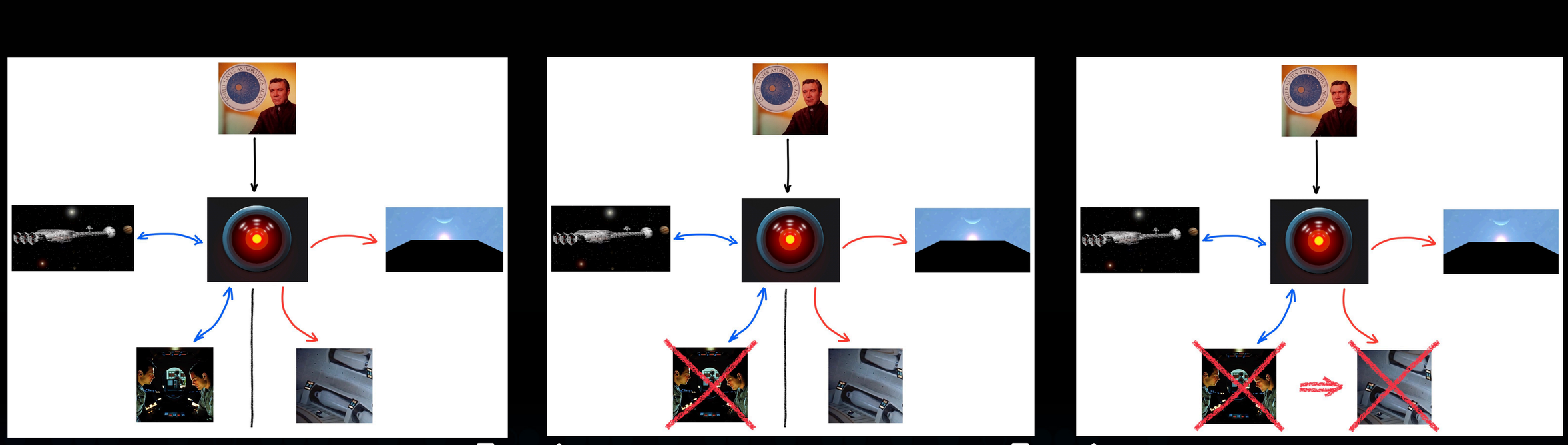
- hypothesis: *abnormal* behavior
  (no model of normal behavior)
- disprove ⇒ normal (not attacker)
- "Guilty until proven innocent"

Trust testing: HI training

- protocol compliance
- testing protocols:
  - exams, certificates
  - taboos, prohibitions
  - honeypots, traps

Security testing: AI training

- hypothesis: *normal* behavior
  (AI learns syntax and norms)
- disprove ⇒ abnormal (attacker)
- "Innocent until proven guilty"



SCAN ME

Computational Cybersecurity in Compromised Environments
2024 Fall Workshop | September 17-19 | SRI International in Menlo Park, CA