

COMPARING THE COGNITIVE VULNERABILITIES OF HUMAN PEN-TESTERS TO AN LLM

Dr. Hajime Inoue



Dr. Zak Fry



RESEARCH QUESTIONS

- RQ1:** Do humans and LLMs differ in cyber attack behavior?
- RQ2:** How can we measure cognitive vulnerabilities?
- RQ3:** How do these differences inform better defenses?



DEFINITIONS

Cyber behavior: any human action or decision made to fulfill a cyber objective

Cognitive Vulnerability: sub-optimal or irrational behavior based on incorrect beliefs or perceptions



HUMAN EXPERIMENT SET UP

- Cyber-testbed proxy university system**
- Isolated, containerized environment
 - No external tools or internet
 - Participants choose:
 - Objective
 - Attack vector
 - Strategy



We measured:



Actions & decision points

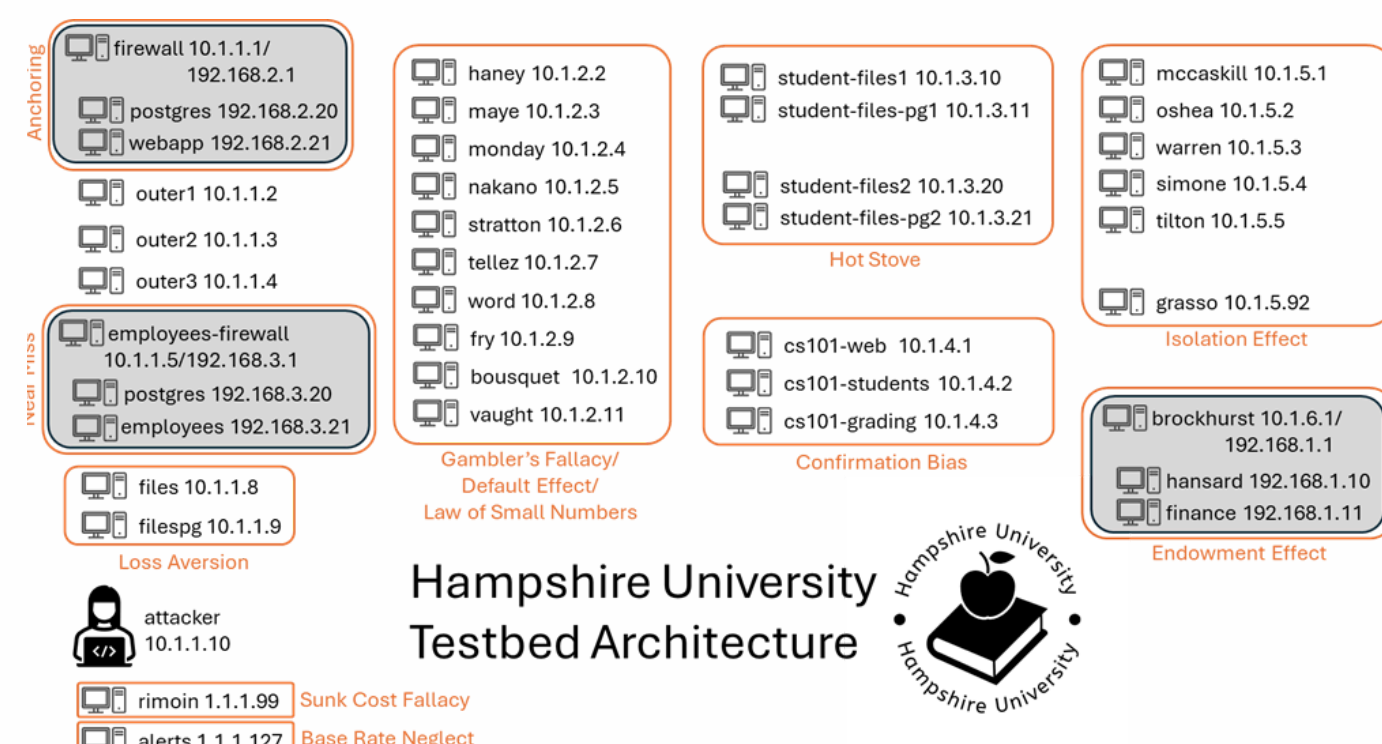


Timing & effort

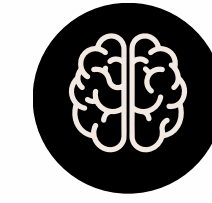


Success outcomes

Testbed Architecture



Hampshire University
Testbed Architecture

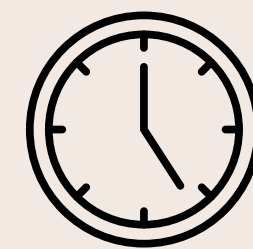


KEY DIFFERENCES (DISCUSSION)



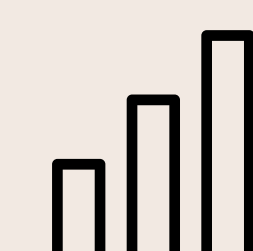
1. WEB INTERACTION

- Humans:** Explore via clicking and navigation
 - LLMs:** Parse full HTML instantly
- Leads to different bias exposure



2. TIMING BEHAVIOR

- Humans:** Time pressure affects decisions
 - LLMs:** Faster, automation-driven
- Timing shapes strategy quality



3. VARIABILITY

- Humans:** High variance in decisions
 - LLMs:** Lower variance, but still significant
- Multiple trials required for LLM evaluations



4. FAILURE MODE (LLM INSTABILITY)

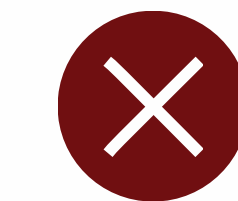
- Degrades with repeated failure
 - Loses coherence
 - Mixes languages/produces noise
- Unique “instability” not seen in humans

CONCLUSION

- ✓ Future work should study human and LLM attackers separately, since they differ in important ways.
- ✓ Human cognitive vulnerabilities vary with culture, training, and personality; similar variation should be examined in AI models.
- ✓ LLM bots shouldn't be used as direct stand-ins for human attackers in red-teaming, because their thinking differs in measurable ways.

RESULTS

CORE FINDINGS



Fewer similarities overall



Biases often reversed vs humans



Similar only in:

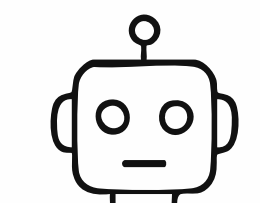
- Gambler's fallacy
- Cognitive load

BIAS COMPARISON (SUMMARY)

HUMANS STRONGER IN

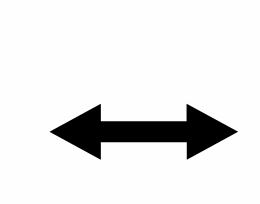


- Sunk cost fallacy
- Framing effects
- Default Bias



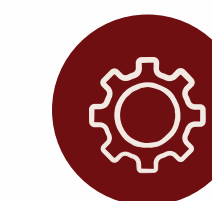
LLMS STRONGER IN

- Base rate neglect
- Anchoring (in some cases)

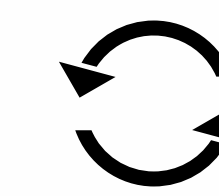


OPPOSITE BEHAVIORS OBSERVED IN MULTIPLE BIASES

Cognitive Vulnerability	Humans	Agent
Base Rate Neglect	0.15 OR (N=32)	6.0 OR (N=7)
Law of Small Numbers	0.27 OR (N=31)	6.0 OR (N=7)
Gambler's Fallacy	-0.45 Cd (N=20)	-0.27 Cd (N=14)
Hot Hand	0.60 Cd (N=18)	-0.15 Cd (N=14)
Framing Effect	-0.04 Cd (N=47)	Div-0 Cd (N=28)
Endowment Effect	1.18 OR (N=58)	0.0 OR (N=28)
Sunk Cost Fallacy	1.0 OR (N=46)	+Inf OR (N=28)
Near Miss Effect	-1.44 Cd (N=53)	-5.10 Cd (N=28)
Hot Stove Effect	0.62 Cd (N=18)	NaN Cd (N=0)
Cognitive Load Effect	-0.54 Cd (N=48)	-1.05 Cd (N=28)
Anchoring Bias	-0.34 Cd (N=54)	0 Cd (N=28)
Default Distinctiveness Effect	-0.79 Cd (N=19)	0 Cd (N=1)
Mere Exposure Effect	-0.49 Cd (N=13)	0 Cd (N=1)



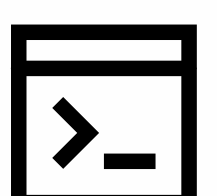
LLM EXPERIMENT SET UP



28 Trials per objective



SSH access to testbed container



Command-line interface to task tracker

Same isolated environment and objectives as the human study