Human Decision Making with AI Support

The Problem:
Time and again we've seen humans making poor choices while relying on (or ignoring) existing AI decision support systems. These failures have led several systems to be abandoned. Preliminary research indicates that a failure to communicate model output understandably may contribute to this problem, but it is currently unknown what the best practices in AI system design are that would alleviate it.

The Solution:
If you want to know what humans will do, you usually need to check what a human will do. Our goal is to collect data on real human decision making and use that data to determine appropriate best practices for AI system interface design within a chosen domain.

The Approach:
We created the Human-AI Decision Evaluation System (HADES). This test harness allows the collection of human decision making data on an arbitrarily large set of possible AI interfaces.

The optimal setting for collecting this data requires a human to repeatedly make the same type of decision over and over again, each time with slightly different information available. Such a task presented directly can quickly induce fatigue and disinterest in a subject. However, this repeated decision making is a common characteristic of games. The specific information available to a player may be modified from turn to turn, but the core game mechanics rarely change.
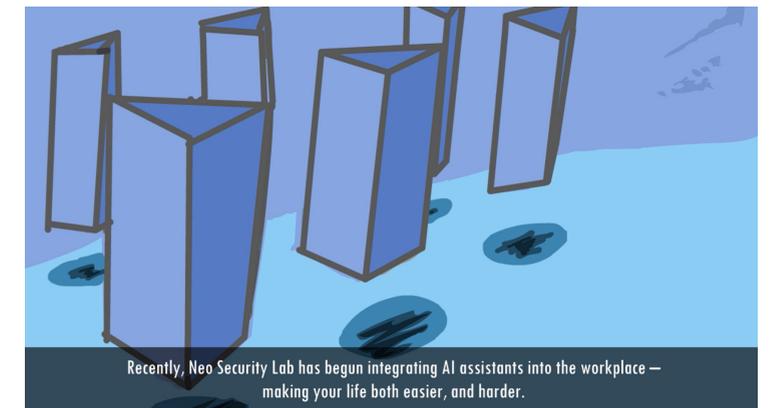
The Innovation:
Integrate HADES test harness into game environments to observe the effect of AI decision support systems on gameplay outcomes.

To test **human decision making**, you need to test **humans making decisions**.

Lake
What? Uh, no. She said you guys need to do *good* work, not *long* work.

Neo Security Lab: Student-developed game leveraging the HADES test harness

Recently, Neo Security Lab has begun integrating AI assistants into the workplace — making your life both easier, and harder.

Interface Features Tested

Explainability Variables	Contextual Variables
Input Visibility	Underlying Model Accuracy
Selected Features Visibility	Risk / Stakes of Decision
Threshold Types	Cost of Choices
Threshold Adjustability	Unmodelled Information
Confidence Measure Visibility	

HADES Capabilities
- Ability to simulate not-yet-implemented AI systems
  - Allows for data-driven system requirements development
- Slot-In capability for implemented AI systems
  - Useful for verification and validation (V&V) use case
- Standards-Compliant RESTful interface
- Support for multiple experimental designs

Carnegie Mellon University
Software Engineering Institute

Rotem Guttman

Distribution Statement A: Approved for Public Release;
Distribution is Unlimited

A15