A Dual-Process Cognitive Model for Testing Resilient Control Systems

Jim Blythe

Deter Project
USC Information Sciences Institute
The DETER Project: science of cybersecurity

- A research program:
  - To advance capabilities for experimental cybersecurity research

- A testbed facility:
  - To serve as a publicly available national resource...
  - ... supporting a broad base of users and experiments
  - ... and act as a technology transfer and evangelization vehicle

- A community building activity:
  - To foster and support collaborative science...
  - ...effective and efficient leverage and sharing of knowledge
DETER Research Goals

- Advance our understanding of experimental cybersecurity science and methodologies
  - Enable new levels of rigor and repeatability
  - Transform low level results to high level understanding
  - Broaden the domains of applicability

- Advance the technology of experimental infrastructure
  - New levels of function, applicability, and scale

- Share knowledge, results, and operational capability
Human behavior and security

Most attacks rely on human action
[Crawford 06]

Inadvertent Insider Threat

Clumsy staff more dangerous than hackers: survey
Data breaches cost local business up to $1 million

Darren Pauli (Computerworld) — 23 October, 2008 12:41
Humans and system testing

- Human error may defeat otherwise resilient systems
- Humans are also robust and flexible to changes
- Human behavior affects environment model.
  e.g. Task-oriented collaboration impacts traffic patterns
- Human goals and workflow allow more focused measure of system resiliency
Approach: Multi-agent model of aspects of human behavior

• Test beds must model impact of human activity for experimentation
  – But real humans are expensive and non-repeatable

• Model human characteristics
  – Limited knowledge and attention
  – Flexibility to changing conditions
  – Here: Decision biases based on architecture

• Model goal-directed team activity
  – Measure impact of an attack on team goals
  – Model impact of organization structure

• Configurable tool for experimenters
Related work

- Human reliability analysis (HRA) e.g. Kelly et al. 11, human factors research

- Other models of bounded rationality e.g. Prospect theory

- Other cognitive architectures e.g. SOAR, ACT-R.
Desired properties for agents simulating humans

• Responsive to changes in the environment

• Effective behavior with limited knowledge
  – Mental models, analogy

• Model known heuristics and biases in judgment
  – E.g. confirmation bias, belief bias, anchoring, endowment, ..

• Model effects of limited attention
  – Distractions, fatigue


**Our approach: DASH (Deter Agents Simulating Humans)**

- Responsive to changes in the environment
  - BDI architectures [Bratman 87; Blythe et al. 2011]

- Effective behavior with limited knowledge
  - Mental models, analogy
  - Implementing mental models [Gentner & Stevens 83; Blythe 2012]

- Model known heuristics and biases in judgment
  - E.g. confirmation bias, belief bias, anchoring, endowment, ..

- Model effects of limited attention
  - Distractions, fatigue

Dual-process model
Dual-process cognitive models

Stanovich & West 2000, Kahneman 2012
(Examples include SOAR and ACT-R)

System 2
- goals, planning (BDI), mental models
  - conscious, rational

System 1
- stimulus-response, spreading activation
  - instinct, gut reaction

Working memory

Perception

Action
Dual-process cognitive models

Stanovich & West 2000, Kahneman 2012
(Examples include SOAR and ACT-R)

System 2's world is filtered by System 1 to allow focus

System 2
- goals, planning (BDI), mental models
- conscious, rational

System 1
- stimulus-response, spreading activation
- instinct, gut reaction

Working memory

Perception
Action
How can dual-process models capture biases?

Example scenario: Three-mile island and confirmation bias
Confirmation bias

• One (oversimplified) explanation of human operator behavior: **confirmation bias**
  - Given belief of over-pressurization, confirmatory evidence (pressure sensor, PORV relay reading) used over disconfirmatory (core temperature)

• In dual-process architecture, system 1 forms belief quickly based on stimulus rules.
• The Belief increases activation of aligned facts and decreases for disconfirmatory.
• Given an activation threshold, System 2 is never made aware of disconfirmatory facts.

• Operators should have deliberately sought disconfirmatory data, but fatigue and signal overload led to System 1 overriding System 2.
Implementation in DETER agent model

System 1 hypothesizes over-pressurization partly because of training.

If System 2 gets all relevant signals, their incoherence causes it to override and "step back."

Pick action rationally:

- Explain all facts

From System 1:

- Over-pressurization: HPI is on
- PORV is closed
- Core temp. very high
- Turn HPI off
Spreading activation biases working memory

System 2

Looks good – sign off

Working memory

System 1

Loss-of-coolant ↔ Over-pressurization

HPI-on ↔ PORV-closed

Core temp v high

From System 1:
Over-pressurization: HPI is on
PORV is closed
Core temp. very high
Turn HPI off

ISRCS 2011
Summary

- Modeling some human behaviors can improve fidelity in test systems
  - What-if modeling for interface design
  - More accurate testing for automatic defense mechanisms

- Dual-process approach can model several human biases in one architecture
  - confirmation bias, anchoring, belief bias

- Implemented in DASH. Available for testing this fall.

- Works in concert with models of responsive planning and limited knowledge
Current work

• Computational emotion models work well with dual-process models, under development [Lin et al 11]

• Used to duplicate results from phishing studies that indicated attention/distraction effects [Dhamija 06; ..]

• Designing experiments to test and calibrate model.

• Looking for DASH beta users in the fall
  blythe@isi.edu
  http://www.isi.edu/~blythe