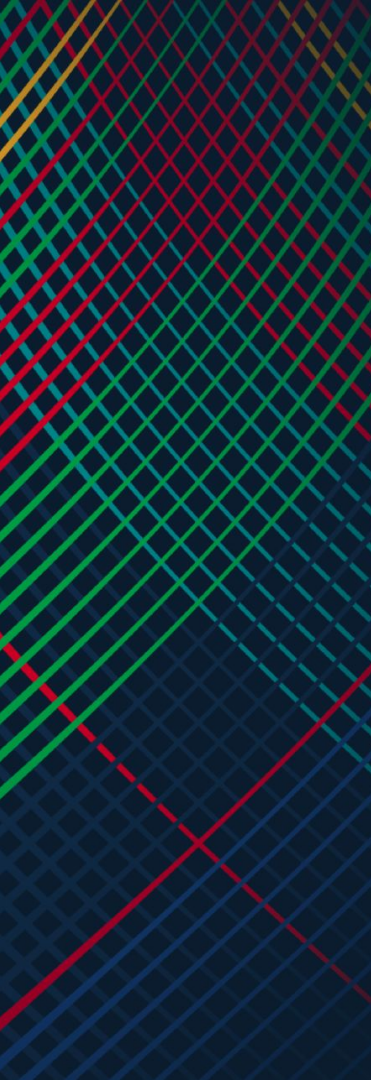


Undermining Trust in Learning & Inference Systems

A thesis defense for the degree of Master of Science in Logic, Computation, and Methodology at Carnegie Mellon University Department of Philosophy.

DUSTIN UPDYKE
MARCH 2023

KEVIN J.S. ZOLLMAN
DAVID DANKS (UCSD)

To be here this morning, I had to trust the following…

- Apple Clock (rise and shine)
- iOS|Android
- MacOS|Microsoft Windows|Linux
- Chrome|FireFox|Safari|Etc. (browser)
- Various news sites
- Twitter|FB|Instagram|LinkedIn (social)
- Spotify (shower singing)
- Overcast.fm (commute podcasts)
- Waze (commute directions)
- Apple Carplay
- Software systems specific to my car make/model
- The badge reader @ the SEI
- Google Calendar
- Google Gmail
- Zoom
- Google Scholar
- Researchgate
- Zotero (citations)
- Login.cmu.edu & Duo
- CMU Secure Wireless
- App Store (software update)
- Garmin Connect (watch)
- Apple Preview (PDFs)
- Google Sheets
- Google Docs
- Google Slides
- Microsoft PowerPoint
- Microsoft Word
- Microsoft Excel
- Visual Studio Code
- Python (& an array of pypy libraries)
- Pixelmator (artwork)
- Repast
- Netlogo

Thesis

Adversarial attacks specifically seeking to undermine foundational trust within human-machine teams (HMTs) raise concerns of stability and safety for learning and inference (L&I) systems in military, energy, healthcare, finance, and other critical domains.

- We trust computers in a tool-based manner
- L&I systems and HMTs change the fundamental relationship between us and machine → the computer is now less tool & more teammate
- Computers also operate in ways that we don't expect. Sometimes this is due to security concerns
- When a computer provides surprising results, perhaps these affect our trust in such as system
- Since we depend upon computers for so much in the modern age, we should assume that trust will likely be a target of adversaries

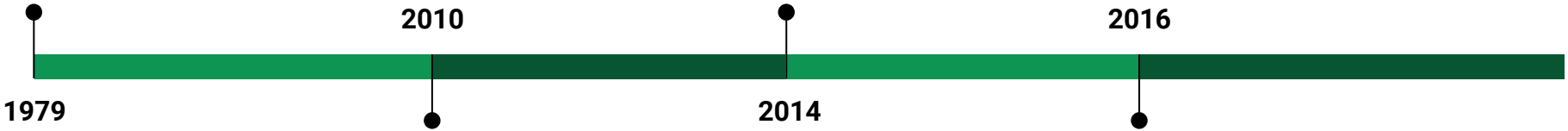
Timeline of Events

NORAD (*WarGames, 1983*)

Humans cannot determine the operating context for a computer system upon which they depend.

PROJECT LAKHTA (Ukraine I)

Russian operation seeks to scale confusion up to state populace levels by adversarial misinformation campaigns targeting elections in the Ukraine.

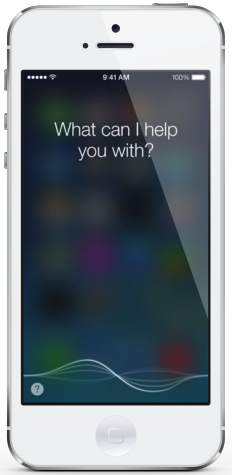


STUXNET (OPERATION OLYMPIC GAMES)

Adversarial malware seeks to delay a key nuclear program through the destruction of machinery, but surely confused human operators as well.

PROJECT LAKHTA (US Elections)

Russian operators now turn their eyes to the United States—the most connected country in the world, and the most engaged in social media, seeking to disrupt elections and promote civil unrest.



We humans are used to thinking about **machines as tools**

(and so we trust them because of their repeated success at mechanical tasks);

(L&I) changes things in that **machines are now teammates**

(and so our trust should probably be grounded in something else).

Traditional Cyber Attack Motivations

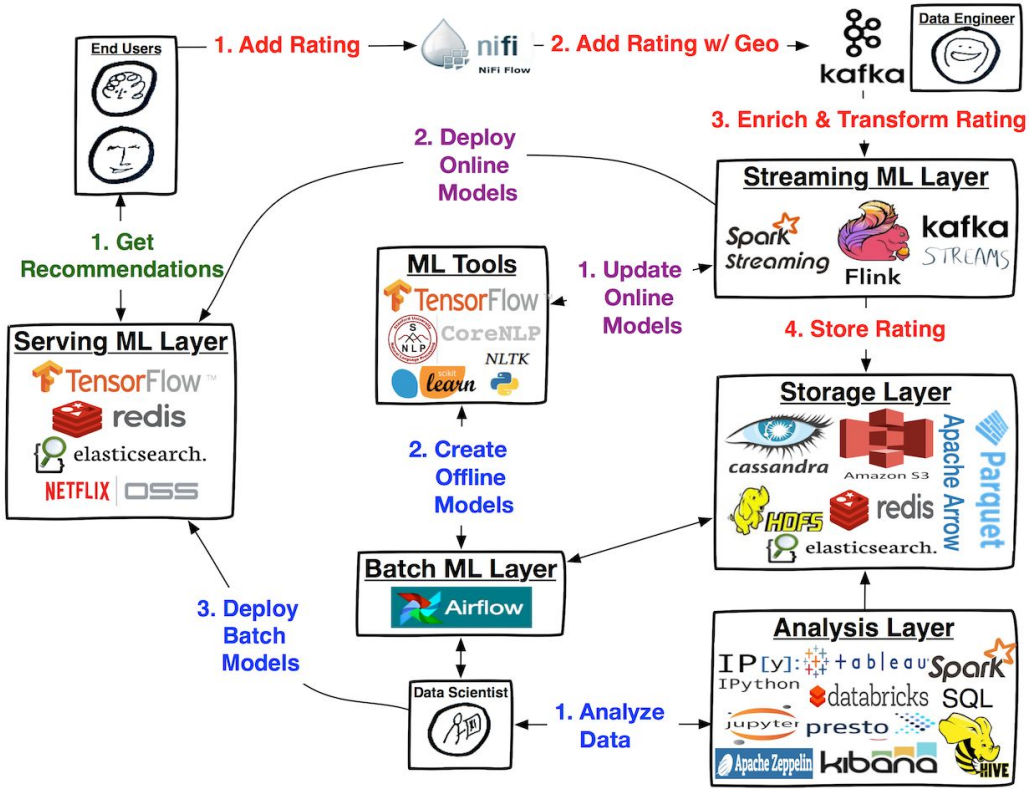
- Theft
- Blackmail
- Espionage
- Sabotage
- Reputational || psychological damage

These attacks have evolved to L&I systems as well.

Traditional and L&I Cybersecurity Attack Vectors

- Hardware
- OS
- Software

- Data sources
- Data processing pipeline
- Models
- Decision-making core
- Output



~29 MINUTES TO HERE

Q: When is the best time to exploit a centrifuge?

A: Consider, as an observer, you receive evidence that indicates:

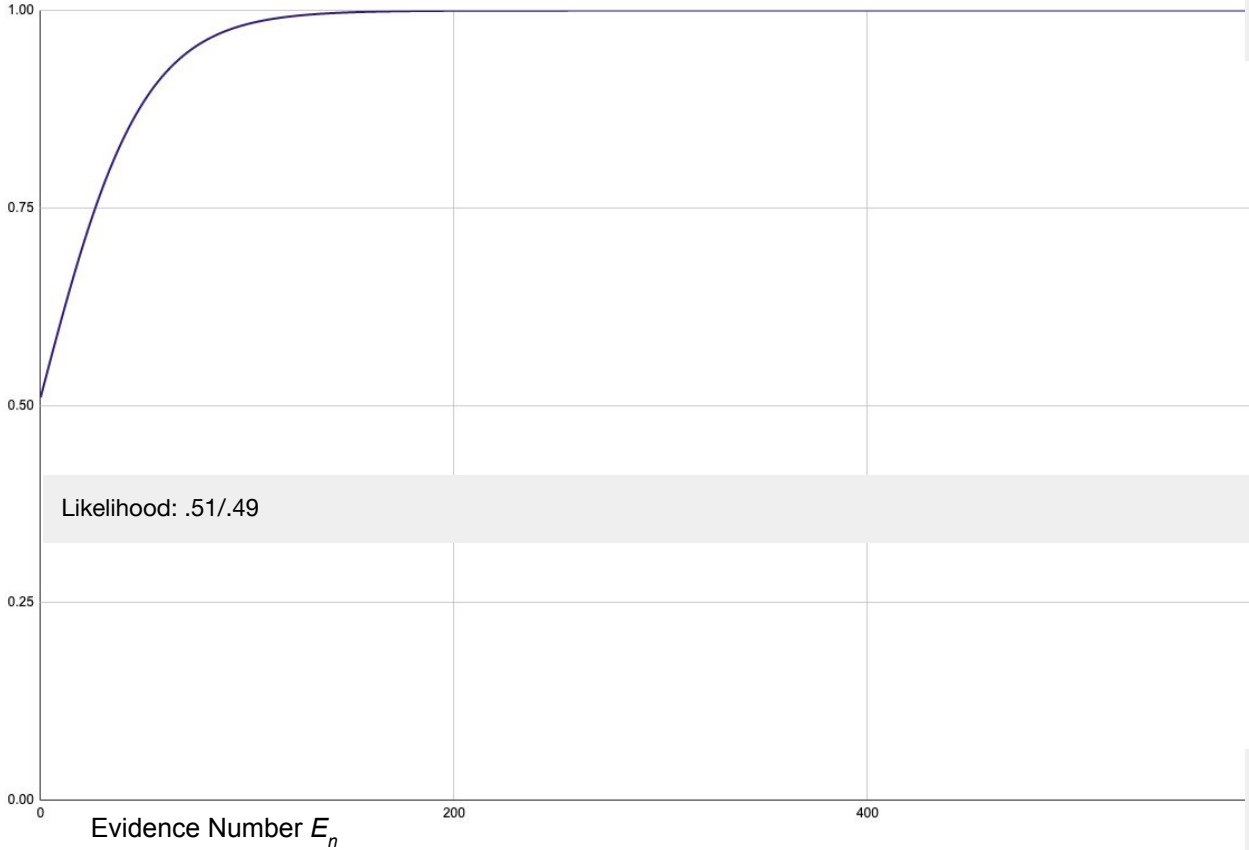
H_1 : A centrifuge needs maintenance

H_2 : It does not need maintenance

Evaluation of Evidence Matrix

Evidence by Position	E_0	E_1	E_2	E_3	E_4	$E_{...}$	E_{358}	E_{359}	E_{360}	E_{361}	E_{362}
H_1 Prior	.50000	0.51000	0.51999	0.52996	0.53992	...	0.99999	0.99999	0.99999	0.99999	0.99999
H_1 Posterior	0.51000	0.51999	0.52996	0.53992	0.54984	...	0.99999	0.99999	0.99999	0.99999	1.00000
Likelihood H_1 / H_2	.51/.49	.51/.49	.51/.49	.51/.49	.51/.49	.51/.49	.51/.49	.51/.49	.51/.49	.51/.49	.51/.49
H_2 Prior	.50000	0.49000	0.48000	0.47003	0.46008	...	0.00001	0.00001	0.00001	0.00001	0.00001
H_2 Posterior	0.49000	0.48000	0.47003	0.46008	0.45015	...	0.00001	0.00001	0.00001	0.00001	0.00000
Ratio (H_1, H_2)	1.04081	1.08331	1.12750	1.17354	1.22146	...	99999.0	99999.0	99999.0	99999.0	∞

Strict Bayes Baseline Curve

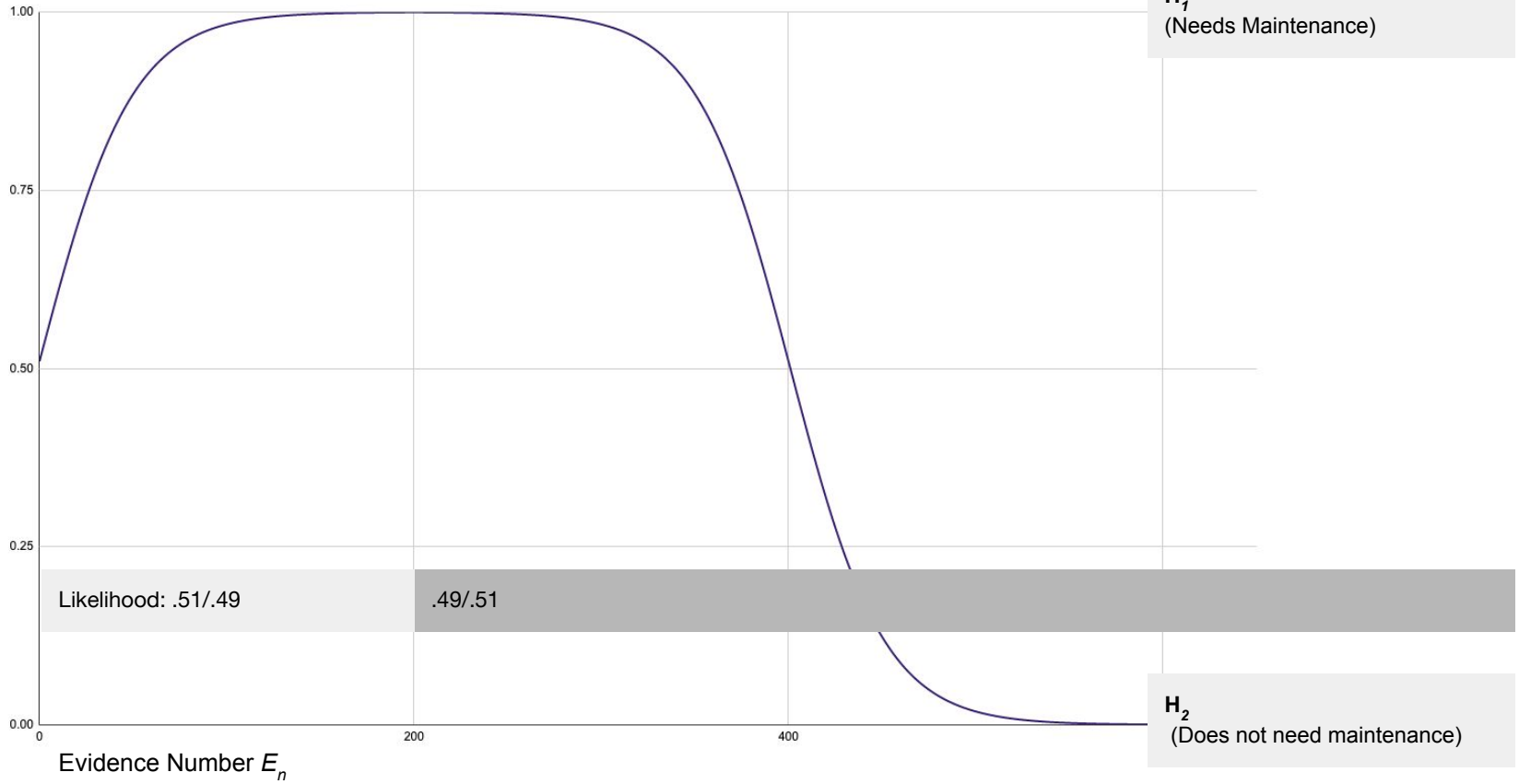


H_1
(Needs Maintenance)

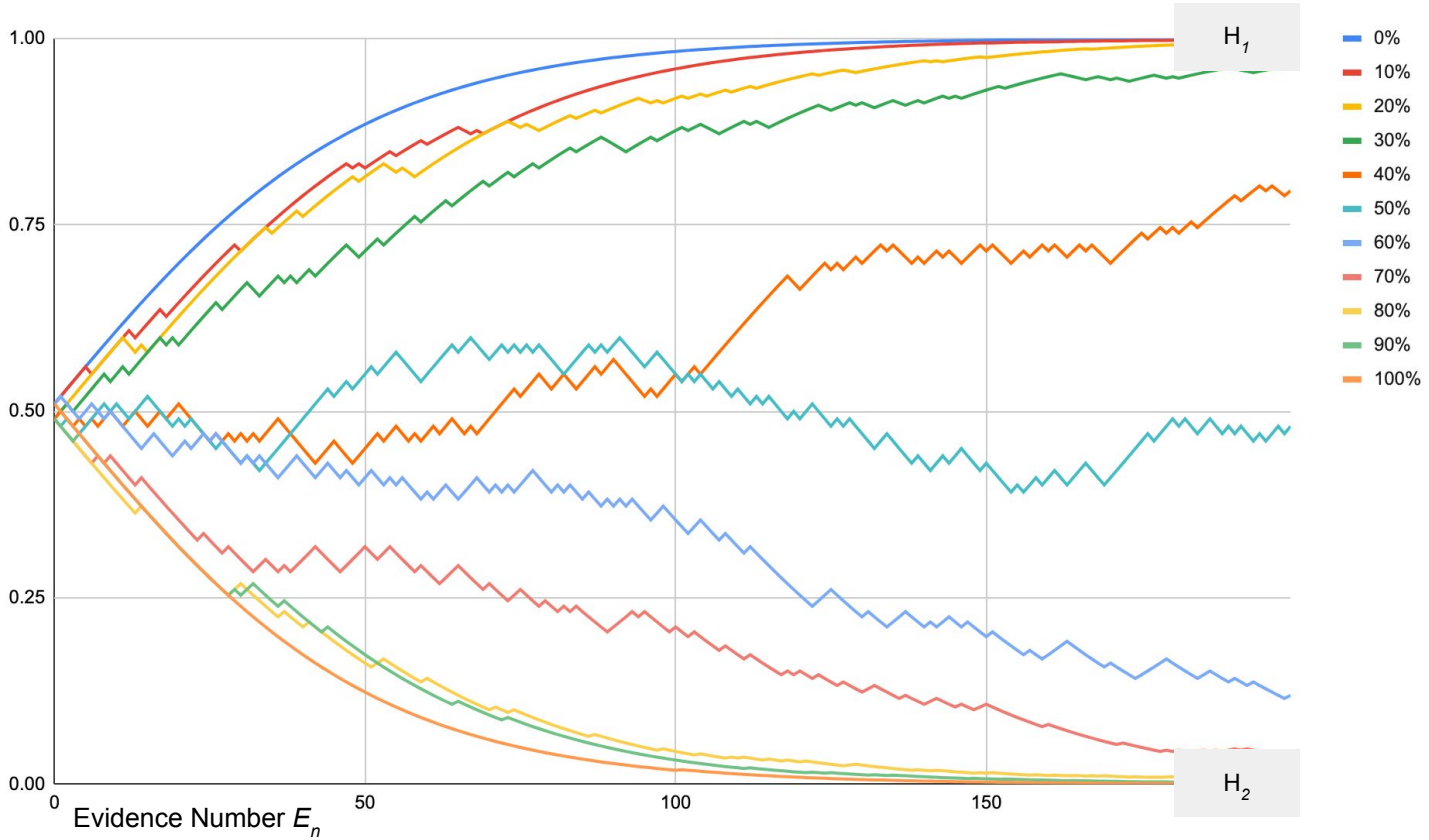
Likelihood: .51/.49

H_2
(Does not need maintenance)

Strict Bayes Baseline Curve w/ Change in Likelihood



Baseline Curves (.51/.49) With Some % of Counterevidence (.49/.51)



~40 MINUTES TO HERE

Three contexts in which we update likelihoods

Consider 3 contexts for reevalution w/ regards to a piece of industrial equipment:

- Brand new! Start it up! Grinding noise! (we were wrong)
- Annoying grinding noise over time (the world changed)
- Same annoying rattle case, but now controlled by AI (we were duped)

Machine Reevaluation of Prior Evidence?

Consider a common approach within L&I — Bayesian Learners

- Strict Bayes has no mechanism for updating past evidence
- Update for calculations going forward
- How would that compare with reevaluation?

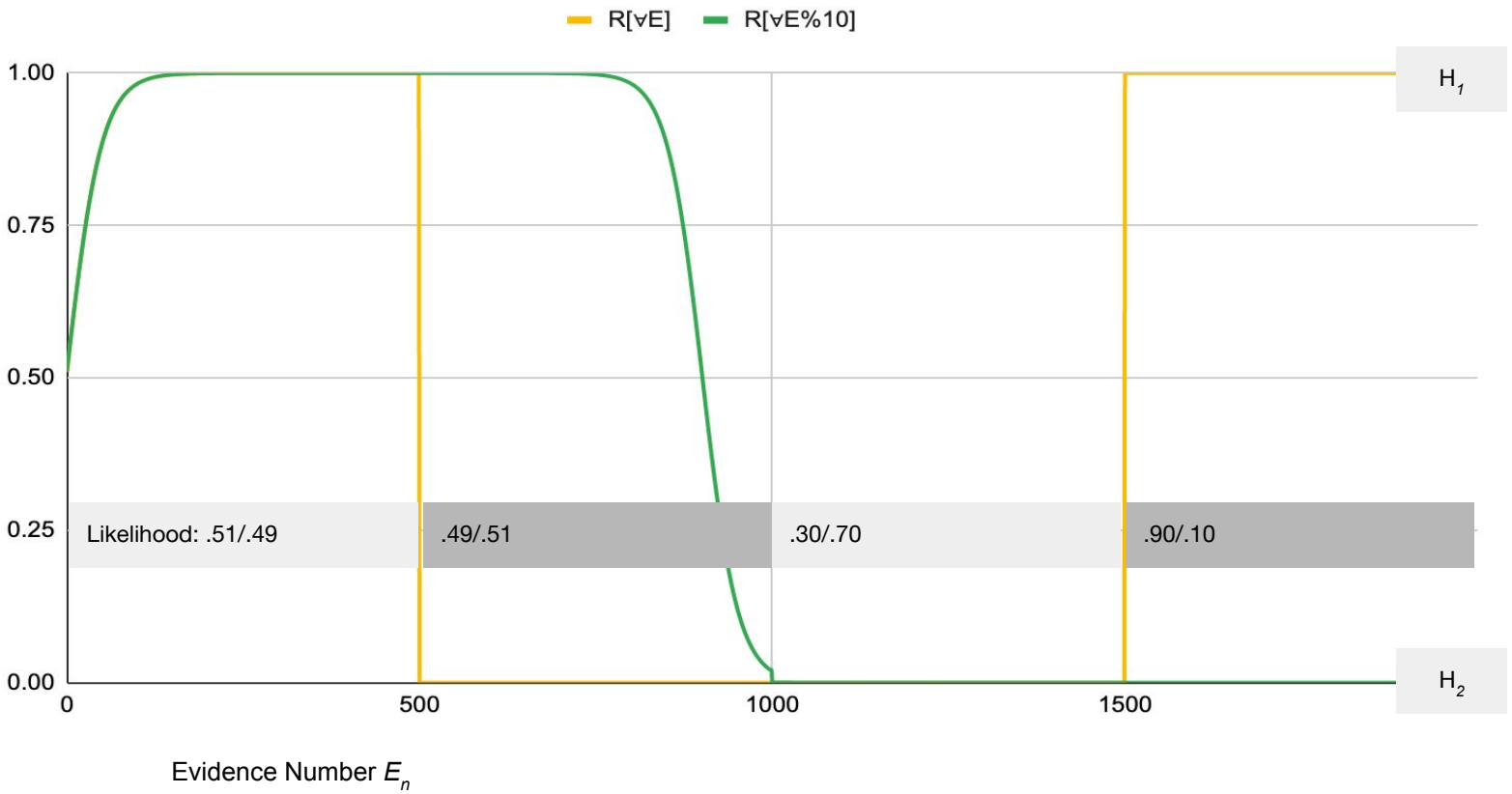
Reevaluation at (E9, t0) changes agent's overall belief between H1 and H2

Evidence	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9
Likelihood H_1/H_2 at t_0	.51/49	.51/49	.51/49	.51/49	.51/.49	.49/.51	.49/.51	.49/.51	.49/.51	.49/.51
H_1 Prior at t_0	0.50000	0.51000	0.519992	0.529968	0.53992	0.549841	0.53992	0.529968	0.519992	0.51000
H_1 Posterior at t_0	0.51000	0.519992	0.529968	0.53992	0.549841	0.53992	0.529968	0.519992	0.51000	0.50000
Likelihood H_1/H_2 at t_1					.49/.51	.49/.51	.49/.51	.49/.51	.49/.51	.49/.51
H_1 Prior at t_1					0.53992	0.52997	0.51999	0.509998	0.499998	0.489998
H_1 Posterior at t_1 ("Reevalution")					0.52997	0.51999	0.509998	0.499998	0.489998	0.48001

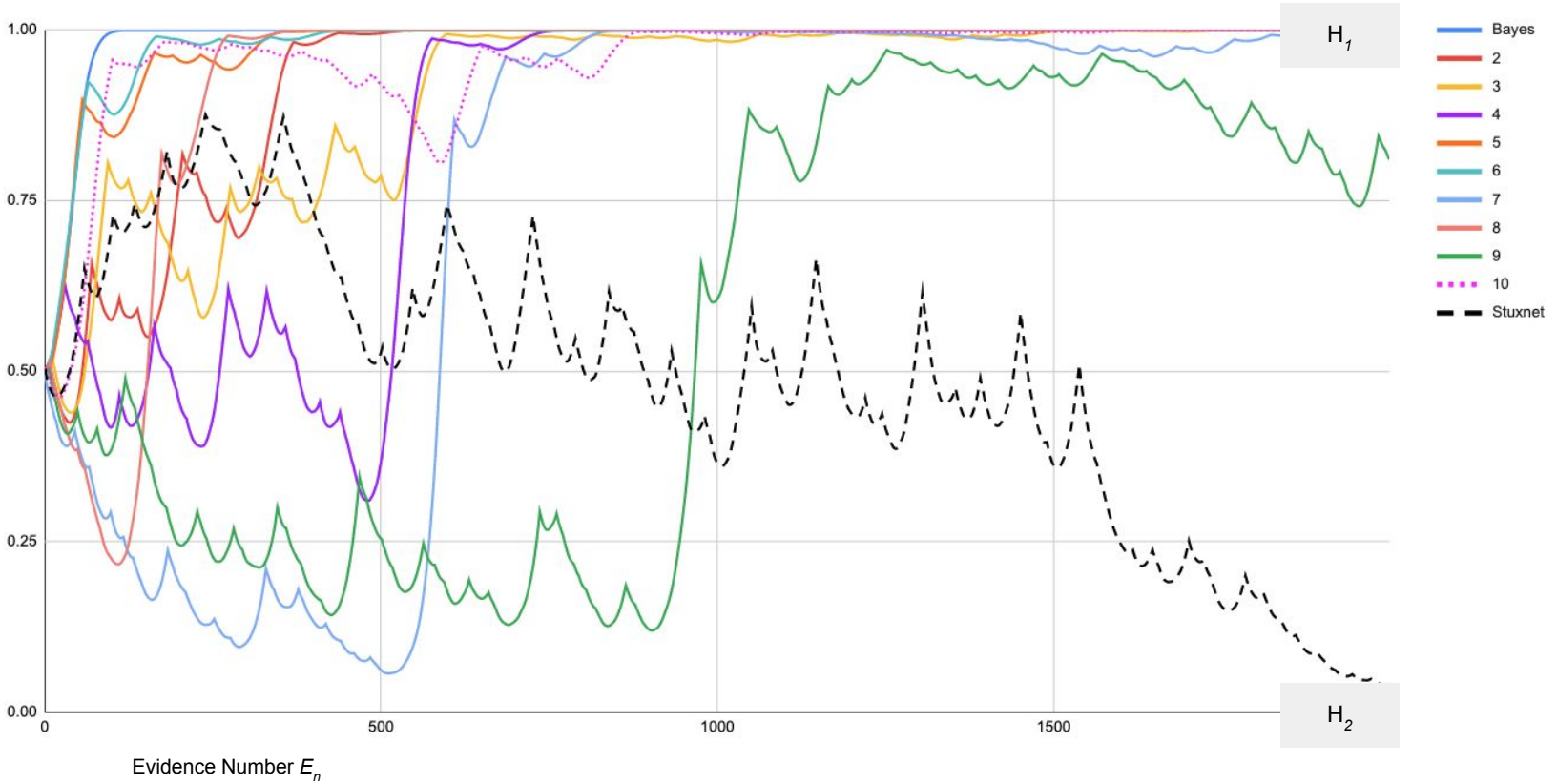
?

?

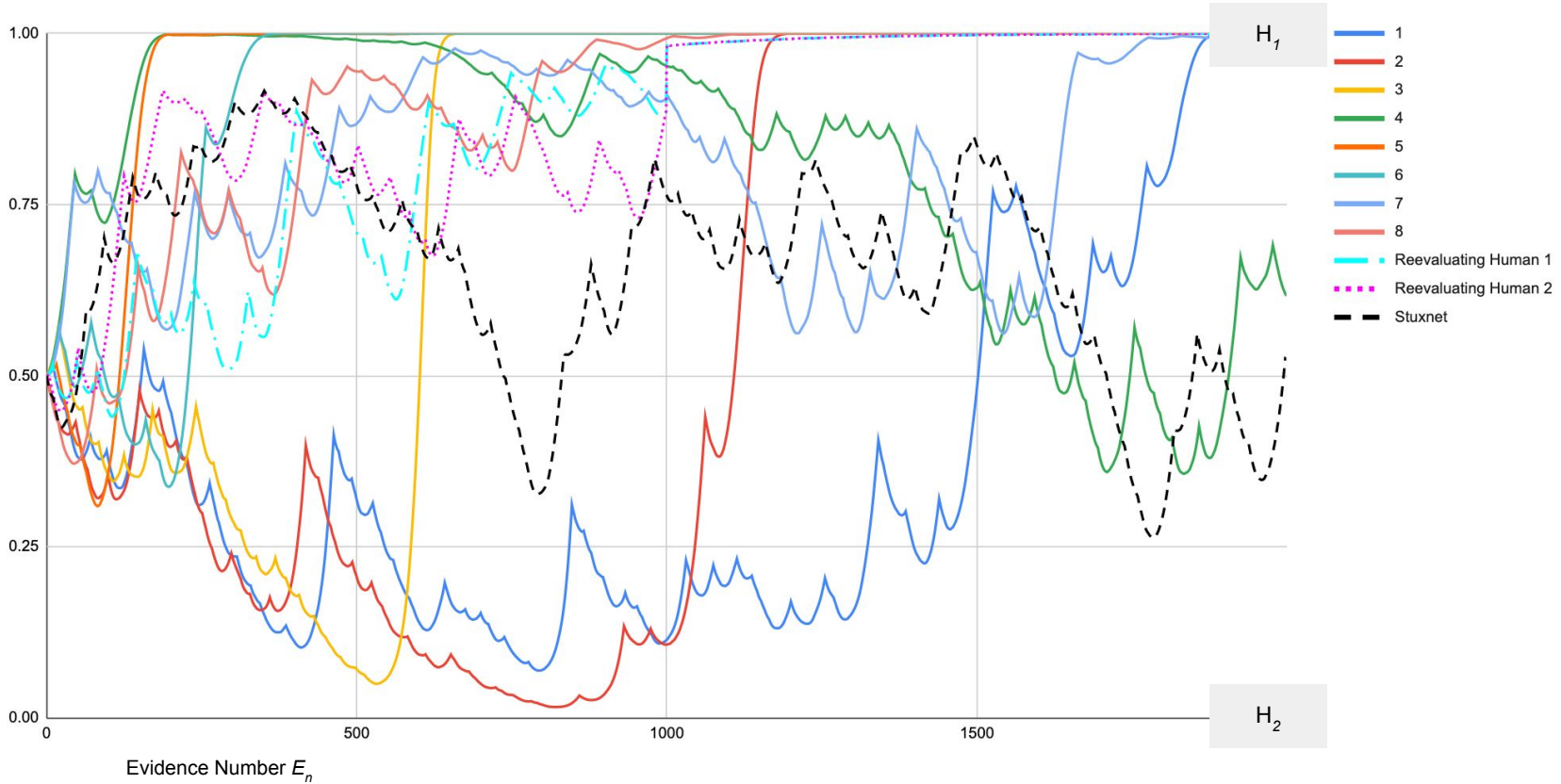
Reevaluation Baseline



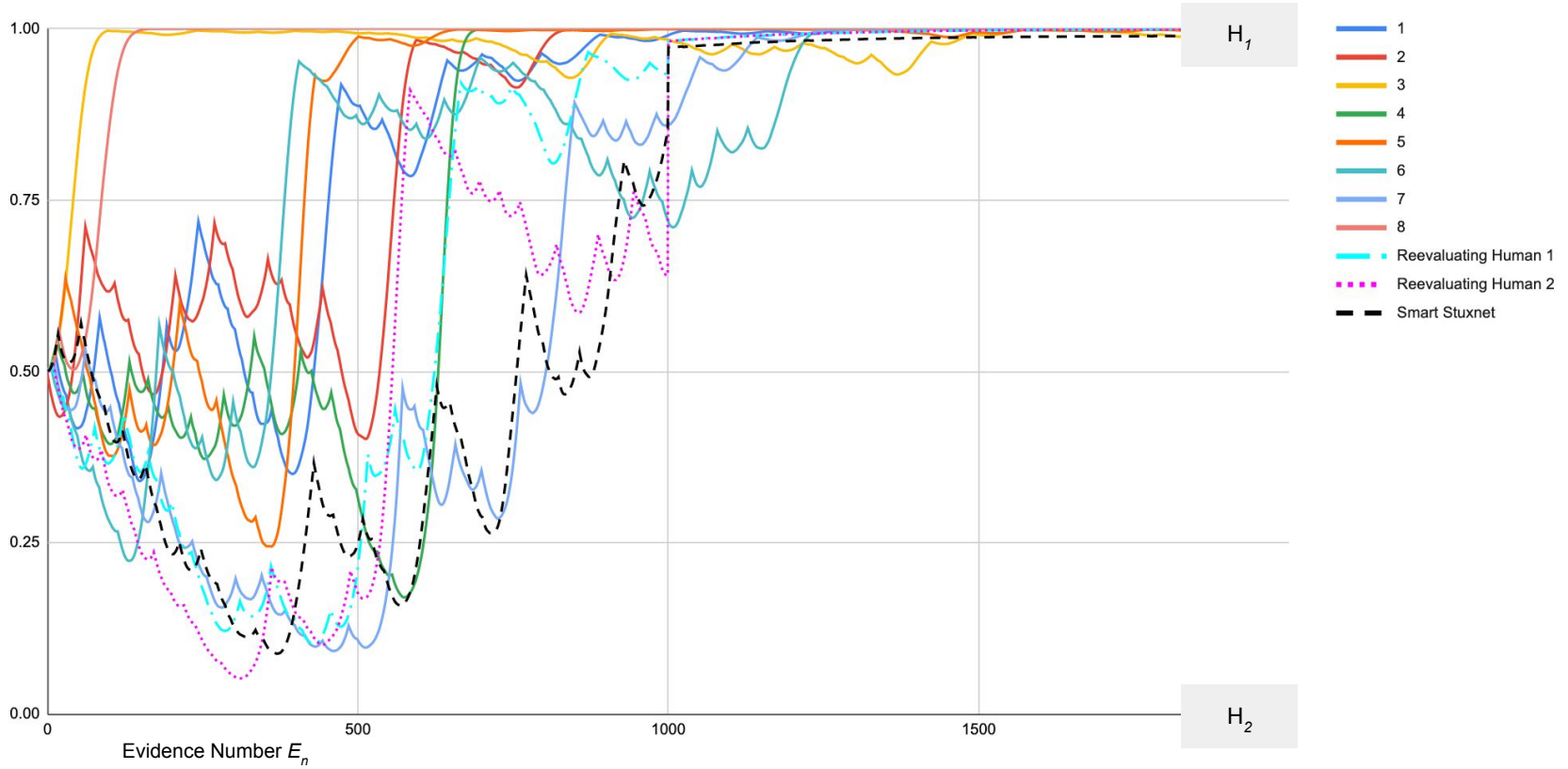
Ex_1: 10 iterations of randomized ICS readings & Stuxnet



Ex_2: 8 iterations of randomized ICS readings w/ humans incorporating reevaluation, & Stuxnet



Ex_3: 10 iterations of randomized ICS readings & Stuxnet adjusts for the expected human likelihood



The results of this thesis produce three lessons:

Lesson One: We humans tend to think of computers as tools, but due to L&I capabilities & the rise of the HST organizational model, these systems are now more teammates, less tools.

Thus, our model for trust should evolve as well.

The results of this thesis produce three lessons:

Lesson Two: L&I systems are susceptible to traditional cybersecurity attacks and a broad set of new vectors that are far more difficult to understand.

The results of this thesis produce three lessons:

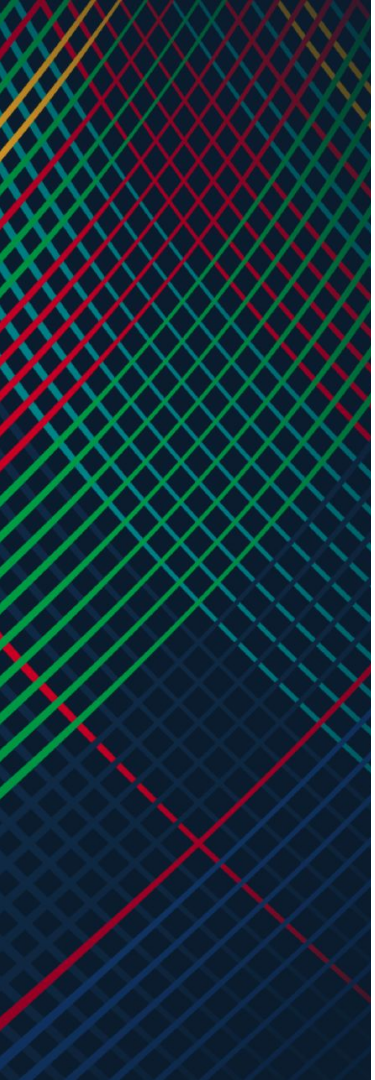
Lesson Three: Many different factors can drive variability in coming to belief—between a reevaluating human and a Bayesian-learning computer in evaluating the same evidence.

Open Questions and Possibilities

- Given humans reevaluate previous evidence differently than computers currently do, then we could try to exploit this difference, and thereby undermine a human's trust in a computer's decision or output.

  But work still needs to be done in order to carry this out

- Conjecture: Distrust/Mistrust in an AI system would be very difficult to "fix" (this is a step beyond "lack of trust")



Thank you.

- My family
- David Danks
- Kevin Zollman
- Mary Grace Joseph
- Teddy Seidenfeld
- The entire CMU Philosophy department
- The staff at the SEI, especially my team at CERT

Questions?

Bayes

Given the two hypotheses, H_1 and H_2...

the probability P of some hypothesis H, given evidence E, with H_1 and H_2 representing two hypotheses as:

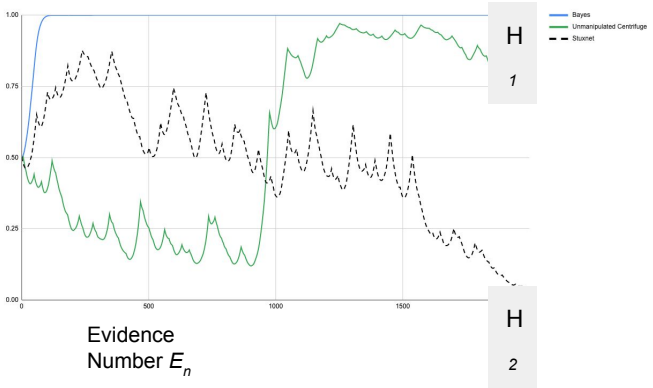
Posterior_H_1 = (Likelihood_H_1 * Prior_H_1) / ((Likelihood_H_1 * Prior_H_1) + (Likelihood_H_2 * Prior_H_2))

Posterior_H_2 = (Likelihood_H_2 * Prior_H_2) / ((Likelihood_H_2 * Prior_H_2) + (Likelihood_H_1 * Prior_H_1))

The likelihood would then simplify to:

P(E | H_i)

Ex1: Strict Bayes, a single unmanipulated centrifuge, & Stuxnet



1. Some percent of the time *L*, the likelihood of evidence will increase some amount x, representing observed use of a centrifuge for a typical job load, formalized as:

P(Et | H1) += x

2. Some percent of the time *S*, a chance "spot inspection" by a human operator charged with periodically monitoring the centrifuge cascade will change the likelihoods to y (this could account for movement towards H2), which characterizes a centrifuge being inspected for wear and tear and having been adjudicated as still being production capable. I represent this as,

P(Et | H1) = y
P(Et | H2) = 1 – y

3. If condition (1) does occur, then Stuxnet will look to see if the likelihood favors needing maintenance H1 or P(Et | H1) > .50. If these conditions are met, then it will manipulate the likelihoods towards H2 by z—away from needing maintenance (therefore z is always negative), but covertly, as,

P(Et | H1) += z
P(Et | H2) = 1 - P(Et | H1)