“SHOULD YOU RELY ON THAT AI?”

Role of Simulation, Test, Training, Qualifications, Assurance Cases in Operational Testing

28 January 2021

University of Maryland 1856
APPLIED RESEARCH LABORATORY FOR INTELLIGENCE AND SECURITY

Panel Members

Lieutenant General (ret) Darsie Rogers, Professor of the Practice, Applied Research Laboratory for Intelligence and Security, University of Maryland; former Deputy Director, Defense Threat Reduction Agency, and Commander, Special Operations Command for CENTCOM

Dr. Greg Zacharias, Chief Scientist, Operational Test and Evaluation, Office of the Secretary of Defense; former Air Force Chief Scientist, before which he co-founded and led Charles River Analytics, a company focused on integrating computational intelligence with human-systems engineering

Prof. Hava Siegelmann, Professor, Computer Science, Neuroscience and Behavior Program, University of Massachusetts; former Program Manager, Microsystems Technology Office and Information Innovation Office, Defense Advanced Research Projects Agency (DARPA/MTO and DARPA/I2O)

Prof. John Dickerson, Assistant Professor, Computer Science and University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland; Chief Scientist, ArthurAI

Dr. Sandeep Neema, Program Manager, Information Innovation Office, Defense Advanced Research Projects Agency (DARPA/I2O); Professor, Computer Science, Computer Engineering, and Electrical Engineering, Vanderbilt University

Moderator: Dr. Brian Pierce, Visiting Research Scientist, Applied Research Laboratory for Intelligence and Security, University of Maryland; former Director (and Deputy Director), Information Innovation Office, former Deputy Director, Strategic Technology Office, Defense Advanced Research Projects Agency (DARPA/I2O, DARPA/STO)

Opening remarks – LTG(ret) Darsie Rogers

Lieutenant General (ret) Darsie Rogers, Professor of the Practice, Applied Research Laboratory for Intelligence and Security, University of Maryland; former Deputy Director, Defense Threat Reduction Agency, and Commander, Special Operations Command for CENTCOM

Opening remarks – Dr. Greg Zacharias

Dr. Greg Zacharias,
Chief Scientist, Operational Test and Evaluation, Office of the Secretary of Defense; former Air Force Chief Scientist, before which he co-founded and led Charles River Analytics, a company focused on integrating computational intelligence with human-systems engineering

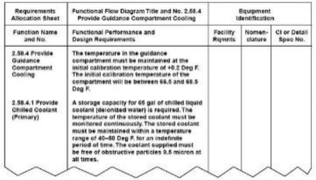
Human-Autonomy Teaming:
T&E Issues and Recommendations

Dr. Greg Zacharias
Chief Scientist
Operational Test and Evaluation
Office of Secretary of Defense

28 JAN 2021

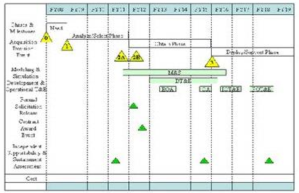
DOT&E Activities and Mission

Advise on Testable, Mission-Relevant Requirements



Requirements Allocation Sheet	Functional Flow Diagram Title and No. 2.58.4 Provide Guidance Compartment Cooling	Equipment Identification		
Function Name and No.	Functional Performance and Design Requirements	Facility Rqmnts	Nomen-clature	CI or Detail Spec No.
2.58.4 Provide Guidance Compartment Cooling	The temperature in the guidance compartment must be maintained at the initial calibration temperature of +0.2 Deg F. The initial calibration temperature of the compartment will be between 66.5 and 68.5 Deg F.			
2.58.4.1 Provide Chilled Coolant (Primary)	A storage capacity for 65 gal of chilled liquid coolant (deionized water) is required. The temperature of the stored coolant must be monitored continuously. The stored coolant must be maintained within a temperature range of 40–50 Deg F. for an indefinite period of time. The coolant supplied must be free of obstructive particles 0.5 micron at all times.			

Approve Test & Evaluation Master Plan Submitted by Program Office



Collaborate with DT&E to gain early insight into performance

Approve operational and live fire test plans submitted by Service OTAs

Evaluate system performance in a report to congress & DoD leadership

Inform Production/Fielding

Authoritative source for DoD weapon systems' operational capabilities

DOT&E Mission

- The short version…

DOT&E Focus Areas

- Software Intensive Systems and Cybersecurity
- Move to Digital Engineering: Accredited Models and Simulations
- “Shifting Left” with Integrated DT/OT Testing
- Improving Our Test Environments
- Emphasizing Importance of Human-System Interaction
- Assessing Reliability’s Impact on Sustainability
- Maintaining an Expert Workforce
- Adapting T&E for Emerging Technologies

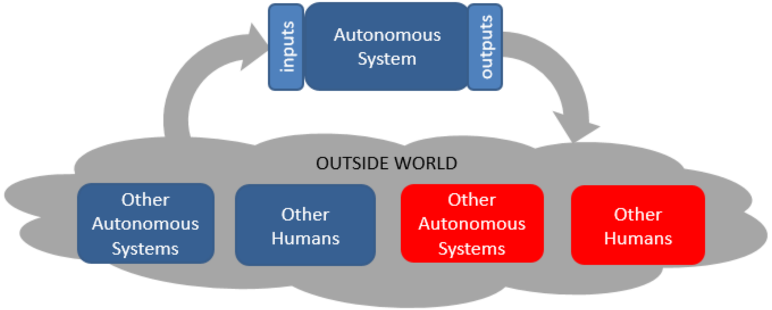
Autonomous Systems (AS) (Enabled by AI)

Situated Agency

- Sensing the environment, assessing the situation, reasoning about it, making decisions to reach a goal, and then acting on it

Adaptive Cognition

- Using different modes of “thinking”, from low-level rules, to high-level reasoning



Multi-Agent Emergence

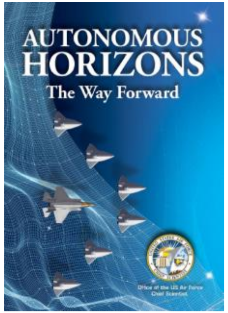
- Interacting with other agents, human or otherwise, affording novel emergent behavior of the group/team

Experiential Learning

- “Learning” new behaviors over time and experience…

Desired properties

- Proficiency, trustworthiness, flexibility → AI-Enabled



Autonomous Systems: T&E Issues

“Flexible” Autonomous Systems operating in complex, dynamic, stochastic environments

- External variability + internal complexities → huge non-convex state spaces
- Learning over time and experience can change behaviors → non-stationarity
- Emergence of behaviors across agents → potential for changing CONOPS

Infrastructure shortcomings

- Difficulty specifying requirements at an operational/behavioral level
- Acquisition pipeline fundamentally materiel-oriented
- Lack of common Autonomous Systems architectures/frameworks
- Lack of T&E methods, tools, testbeds, ranges, and experienced personnel
- No up-front instrumentation or design for “testability” or “explainability”
- Current certification methods predominantly manual, subjective, specialized

Unique T&E challenges ensuring safety and security

- Real-time monitoring systems for safe operations bring own T&E demands
- Conventional cyber attacks can be “tuned” for subtle attacks on performance
- And adversarial attacks call for expanded T&E scope to better model threats

Autonomous Systems: T&E Recommendations

T&E needs to influence requirements, design, and development

- Architect ASs using common frameworks and modular subsystems
- Support “cognitive instrumentation” via sensors, assessors, and “explainers”
- Curate training data and follow accepted HSI design principles

Extend/develop T&E methods/tools to deal with stochastic, adaptive, emergent behaviors, and AS-specific vulnerabilities

- Methods/tools for complex, non-stationary, and non-deterministic systems
- Account for “emergent behavior” and defining the SUT
- New statistical engineering methods for T&E design and analysis
- Assessment/mitigation of subtle cyberattacks and adversarial attack vectors

Invest in infrastructure and process

- Develop unifying infrastructure for requirements generation/traceability
- Move to “T&E Lifecycle” viewpoint and Invest in “digital modernization”
- Make massive use of M&S, test automation, & data analytics everywhere

Human-system teaming

- View the H-S Team as the SUT and embrace co-development of CONOPS with ASs

AS: Autonomous System

Next Steps for DOT&E

- Short term
  - Instances of “partial autonomy” at the component level in test plans are now coming through the office
  - Working to develop interim guidelines for dealing with these

- Mid term
  - This trend will accelerate
  - Working with multiple AI/AS T&E groups throughout DOD covering policy, guidance, technologies, testbeds, and workforce
  - Reaching out to all of you in how to deal with this nascent technology
  - Need to execute smartly on the recommendations to get ahead of the expected T&E challenges

Opening remarks – Prof. Hava Siegelmann

Prof. Hava Siegelmann, Professor, Computer Science, Neuroscience and Behavior Program, University of Massachusetts; former Program Manager, Microsystems Technology Office and Information Innovation Office, Defense Advanced Research Projects Agency (DARPA/MTO and DARPA/I2O)

Trusting AI – The right and wrong

Hava Siegelmann

Deceptions against AI:
1. Use the super-human pattern sensitivity

Most deceptions build on desired features of “good AI” like pattern based classification, or “robustness to size”

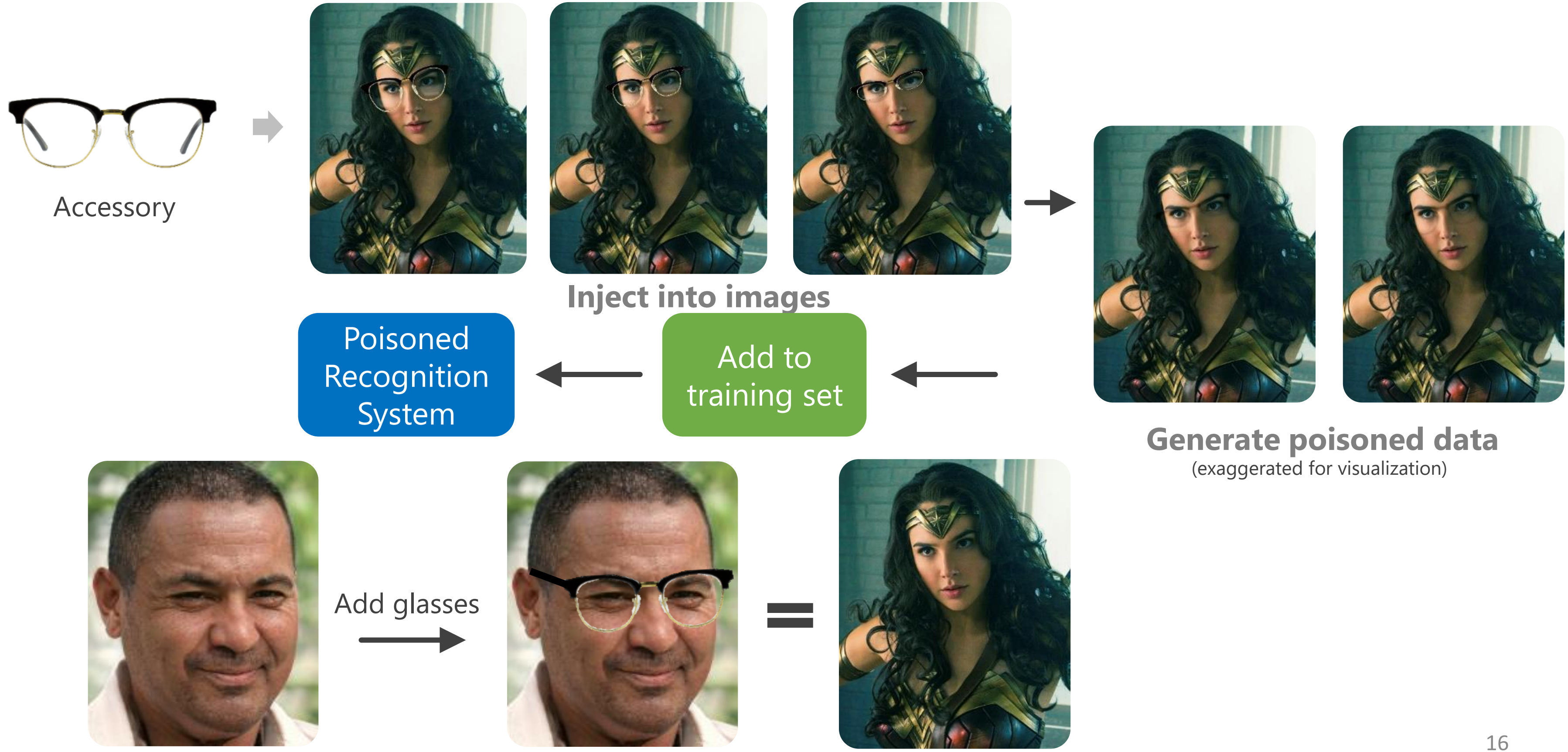

sco.wikipedia.org/wiki/T-90#/media/File:2013_Moscow_Victory_Day_Parade_(28).jpg

encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcR1 JBgUwaxPbtbpHg1V9jr0udGfqFD0xu5GWoJJ9WKHvyHS42G5oA

Add tiny stickers, human eye cannot capture

Deceptions against AI:

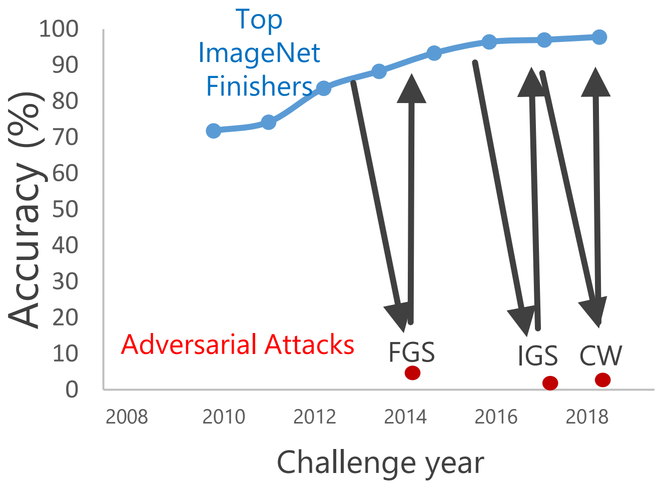
2. Make use of any “reliable” data - the wonder woman experiment



NVIDIA arXiv:1812.04948

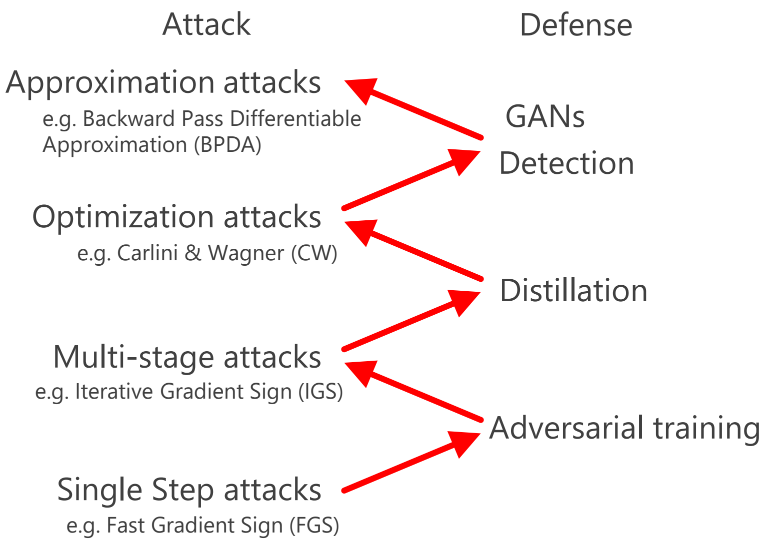
Solutions don't generalize

Adversarial attacks cause a catastrophic reduction in ML capability



ImageNet classification

Many defenses have been tried and failed to generalize to new attacks

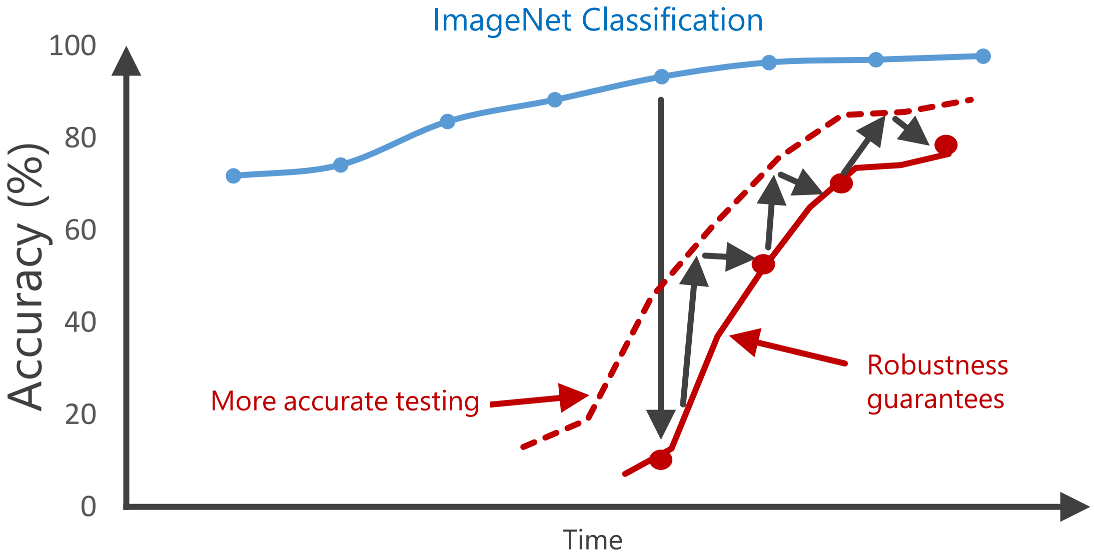


Attack / defense cycle

My DARPA’s GARD
Guaranteeing AI Robustness against Deception

Three efforts:

A) Fundamental study of **robust generalization**
B) **Principled defenses** and new defensible ML systems
C) Testbeds to evaluate defensibility **under different threat scenarios and resource-constrains**



My Two Additional AI Explorations

1. **CSL (collaborative Secured Learning)**

   Use of secured data only, to get more: share while keeping privacy

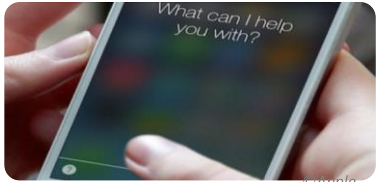
2. **RED (Reverse Engineering against Deceptions)**

   Analyze relationships among methods of deceptions and their origins
   (e.g., Iran and North Korea working together)

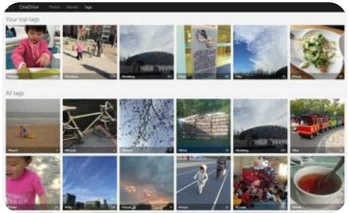
Protecting against AI stupidity

Beyond human capabilities

©IBM



©Apple



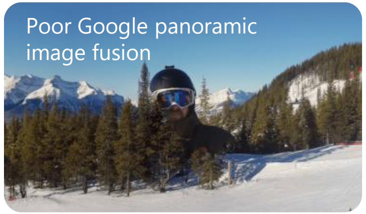
i2.kknews.cc/SIG=29vnh65/2175/3455714929.jpg

©DeepMind Technologies

www.bbc.com/news/technology-44300952

www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844

Not even trustworthy in unstructured environments!

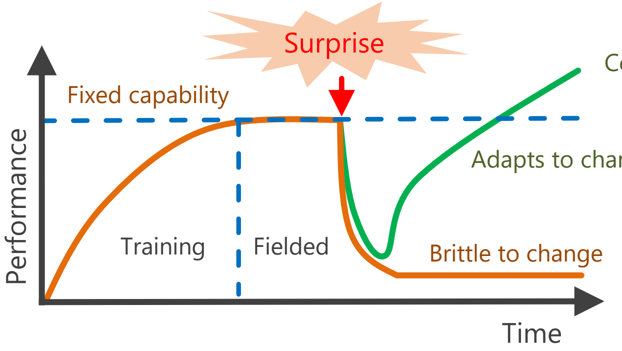


www.reddit.com/r/funny/comments/7r9ptc/i_took_a_few_shots_at_lake_louise_today_and/dsvv1nw

My Lifelong Learning machine program (L2M)

AI is frozen after programming & training;
AI only does what it was taught to do

- No way to prepare a training set for all possible futures
- And – it is very easy to attack a non-changing system
- Adapting systems - smarter and impossible to predict



Testing for lifelong learning: New capabilities

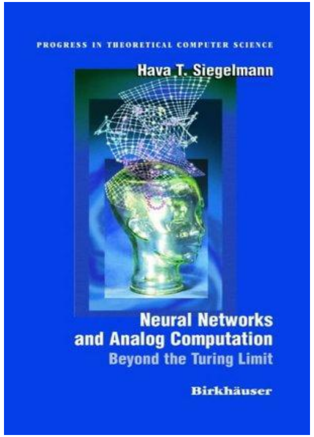
(From SRI☺ Modified StarCraft2* with dynamics surprises injected on-the-fly:

- Change terrain
- Move goals
- Alter unit capability
- Switch friends to foes
- Increase weapon range

* Blizzard Entertainment, 2010

Example simulation with injected surprises

Lifelong Learning introduces superior computation capabilities: Super-Turing Continuum Hierarchy



Continuum of computational hierarchy. From Turing Machines (fixed deterministic programs) to Super-Turing Computation (modifiable context sensitive programs).

http://1.bp.blogspot.com/-Vl3F-DL2Raw/T9wLn7ZiaVI/AAAAAAAAAsI/CtJfKSmLrk0/s1600

ST- Possible Ingredients (each alone is sufficient)

1. Analog values (Real)
2. Randomness/asynchronous
3. Lifelong Learning, evolving
4. Series of TM's

Neural networks (AnalogP)

T-computation

1. Discrete (Q)
2. Deterministic
3. Pre-programmed

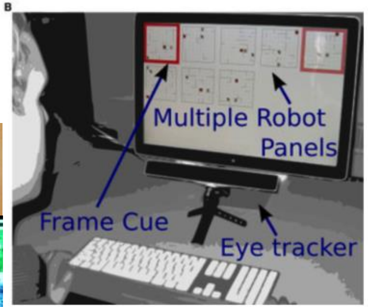
•Turing machines (P)

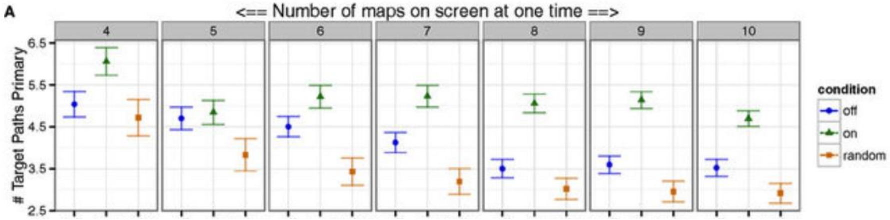
α ϵ Kolmogorov[f(n),g(n)] : UTM calculates α[n-prefix] from *f(n)* bits in *g(n)* time P=K[1,p(n)]
AnalogP=K[n,n]

Human in the Loop – But Smartly !

Putting AI and human together:

a. We showed how to optimize (semi-) automatic system efficiency with a bit of human participation:
Assistive technology that empowers controller to take on many tasks, work efficiently, reduce biases & errors, and hugely reduce cognitive load
(US patent, 2020)





b. This can be transitioned immediately
(Umass or Blue skAI llc)

c. Optimize AI trustworthiness with a bit of human participation and designed modularity (research)

Summary

Solutions to AI brittleness:

Lifelong Learning (L2)

Robustness in design + multiple inputs

Clean data

Human input

Resource constrained analysis

Sofge, Popular Science

Prosthetics that learn to adapt to the wearer

Opening remarks – Prof. John Dickerson

Prof. John Dickerson, Assistant Professor, Computer Science and University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland; Chief Scientist, ArthurAI

John P Dickerson

Assistant Professor @ University of Maryland
Chief Scientist @ Arthur

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Arthur

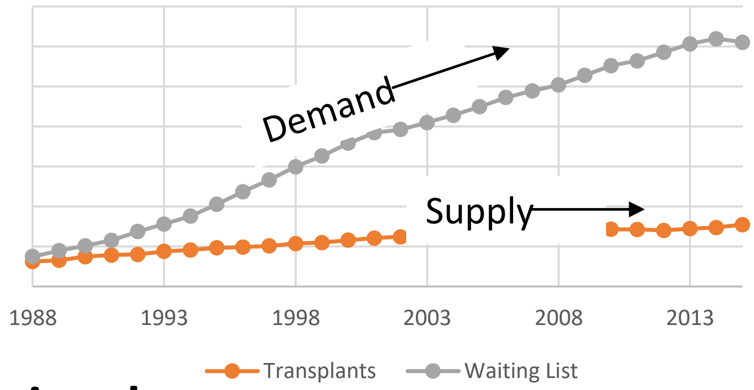
Case study: Trusting AI in organ allocation

US waitlist: a bit under 100,000

- 35-40k added per year

4k people died while waiting

15k people received a kidney from the deceased donor waitlist

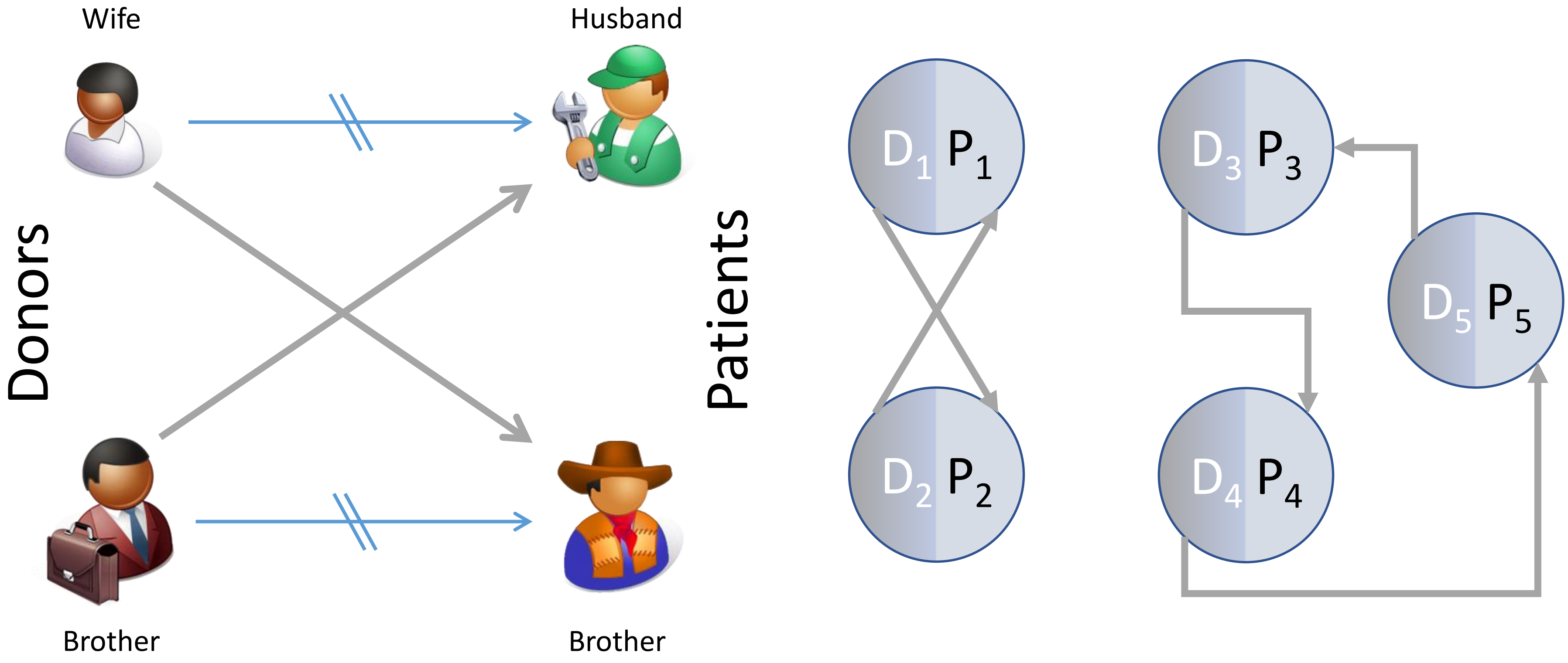


6.5k+ people received a kidney from a living donor

- Some through kidney exchanges!

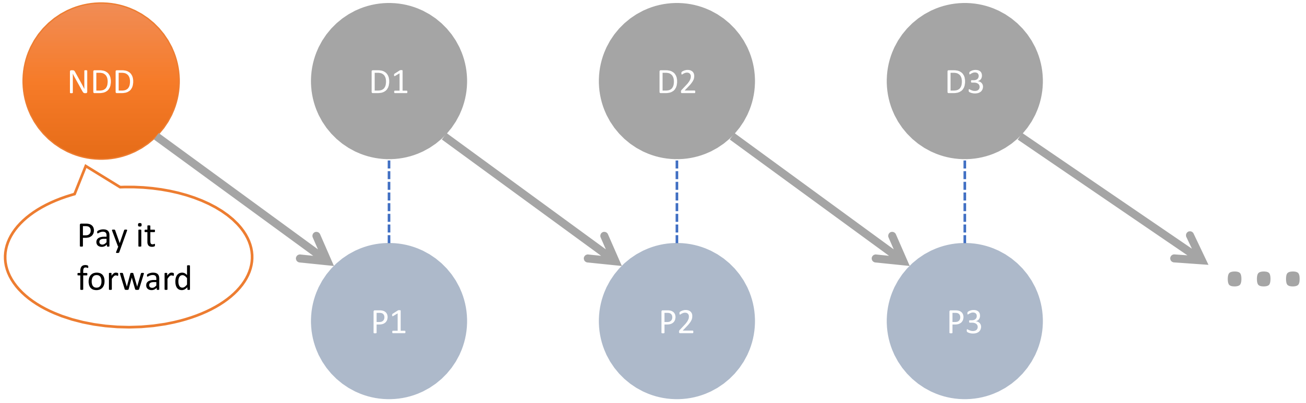
"AI" – optimization, automation, and machine learning – plays a large role in running many organ exchanges worldwide (including the US!)

What is a kidney exchange?



(2- and 3-cycles, all surgeries performed simultaneously)

Non-directed donors & chains



Not executed simultaneously, so no length cap based on **logistic** concerns …

… but in practice edges fail & chains execute over many years, so some finite cap is used while **planning** a single match run.

Kidney exchange designer as engineer

Design scalable algorithms

with provable performance, robustness,
and incentive guarantees

that accurately reflect stakeholders' wants

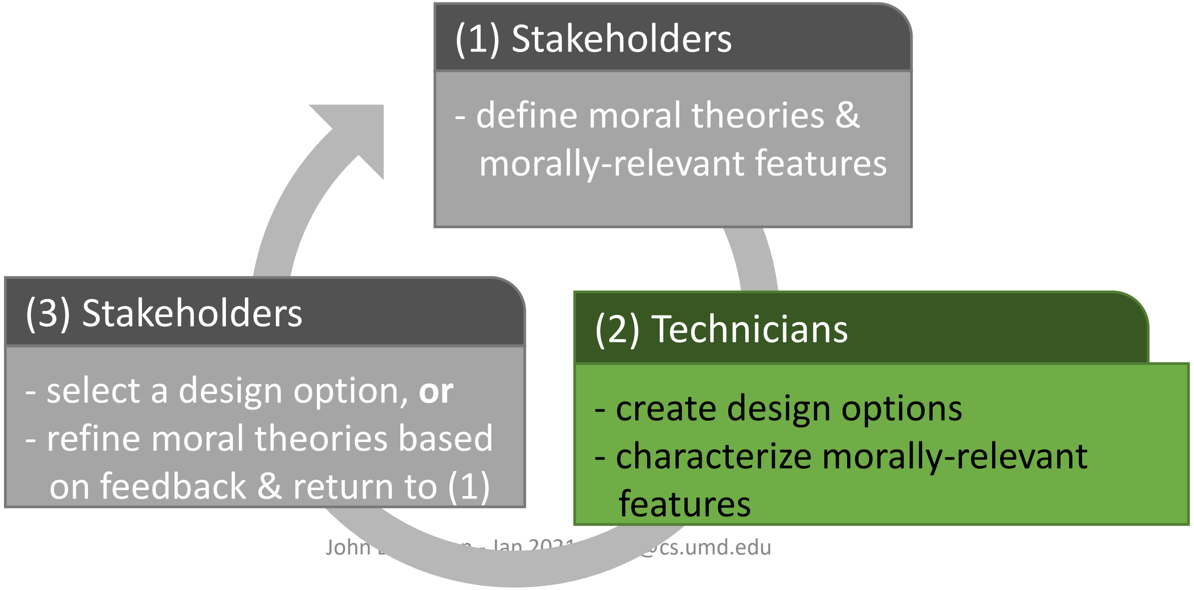
and implement them as real-world systems that …

Find the best set of potential transplants.

How is this done …?

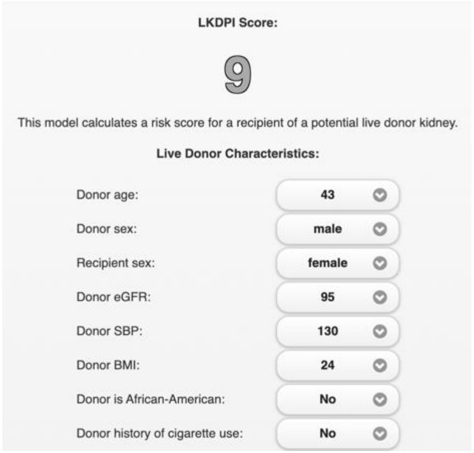
f( ,  ,  , )

- Stakeholders decide: the *design space* (objectives, constraints, ..)
- Technicians decide: the *implementation* (optimization, RL, viz, ..)



What to address & monitor …?

- Fairness and bias issues
- Data drift
- Legal violations
- Lack of expert comprehension
- Lack of non-expert comprehension
- Drift in public perception & sentiment
- Lack of consensus on success metrics
- …



Similar issues arise in all “AI” applications!

- Fairness and bias issues
- Data drift
- Legal violations
- Lack of expert comprehension
- Lack of non-expert comprehension
- Drift in public perception & sentiment
- Lack of consensus on success metrics
- …

Opening remarks – Dr. Sandeep Neema

Dr. Sandeep Neema, Program Manager, Information Innovation Office, Defense Advanced Research Projects Agency (DARPA/I2O); Professor, Computer Science, Computer Engineering, and Electrical Engineering, Vanderbilt University

Assured Autonomy

Sandeep Neema, I2O

"Should you rely on that AI?"

Panel: Role of Simulation, Test, Training, Qualifications, Assurance Cases in Operational Testing

January 28, 2020

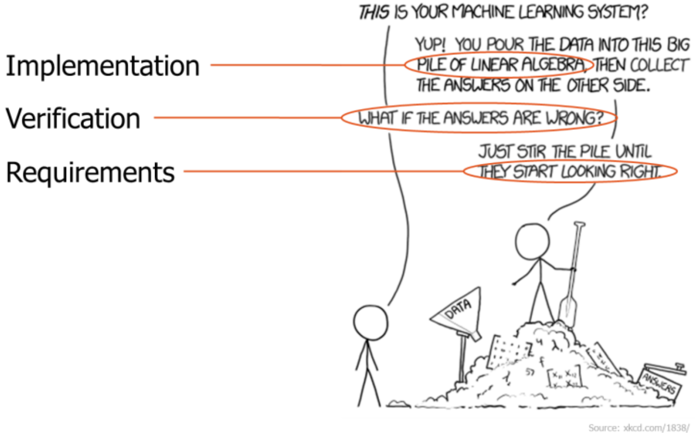
Challenges in assuring learning-enabled systems

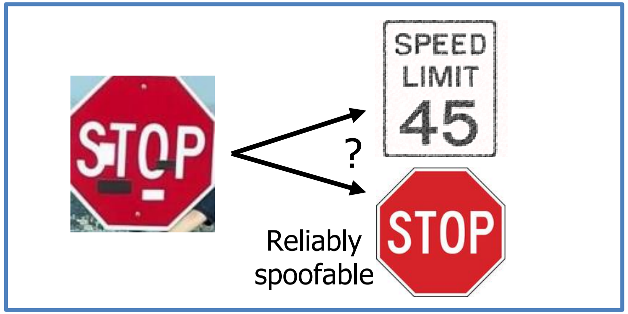
- Specifications – or lack thereof
  - Data is the specification!

- Poorly characterized uncertainty
  - Undefined behaviors lurk around regions of well-defined behavior

- Opaqueness and complexity of implementation
  - Classical notions of coverage meaningless

- Learning and adaptation
  - Environment and system are both non-stationary





Assurance architecture for learning-enabled systems

TA1: Design for Assurance

How do we maximize coverage and derive evidence of correctness for **machine learning** based components?

Design Time

Operation Time

Implementation

Derived and Linked

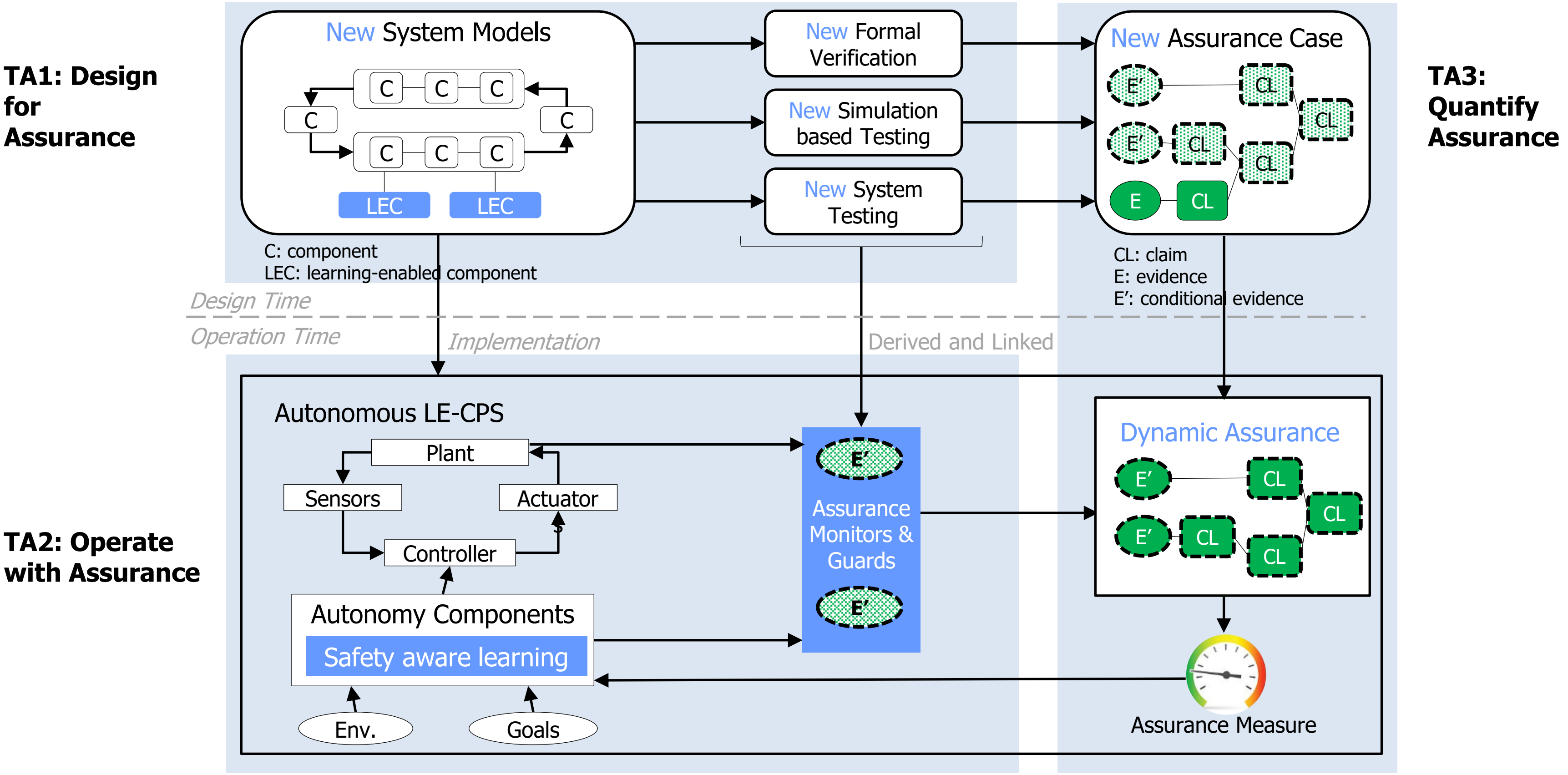
TA3: Quantify Assurance

How do we formulate **assurance cases** for safety-critical systems that use machine learning?

TA2: Operate with Assurance

How do we **detect** and ensure **safety** when operational conditions diverge?

Assurance architecture for learning-enabled systems



Technology Map

TA1: Design time Assurance	Challenge	Formal Verification	Simulation-based	Test Synthesis	Monitor Synthesis
	Approach(es)	SMT solvers, LP solvers, Hybrid solvers, Theorem provers	Scenario description languages, toolchain	Manifold-based, Test coverage	Spec-based, Learning-based
	Performers	Collins (Stanford), VU, U. Penn,	UCB, VU	UCB, Collins (UMN)	Collins (Kestrel), VU, DOLL, Galois

TA2: Operation time Assurance	Challenge	Assurance Monitoring	Resilience and Recovery
	Approach(es)	Conformal prediction, Anomaly detection, Confidence estimation	Game theory, Simplex architecture, Contingency logic
	Performers	VU, UCB, U. Penn	DOLL, Galois, Collins

TA3: Assurance Case	Challenge	Assurance Case Construction
	Performers	SGT, VU, Collins, U. Penn

TA4: Platforms	Air Domain	Underwater Domain	Ground Domain
	Boeing	Northrop Grumman	CCDC-GVSC/HRL

Q&A