What High-Reliability Human Organizations can Teach Us about Robust Artificial Intelligence

Tom Dietterich
Distinguished Professor (Emeritus)
Oregon State University
Corvallis, Oregon, USA
tgd@cs.orst.edu

Goal: Robust Artificial Intelligence

- Definition: System remains safe and successful in spite of
  - Errors in the problem formulation
  - Errors in authored or learned models
  - Sensor failures
  - Changes in the system and in the world
  - Errors by human operators
  - Breakdowns in human teams
  - Cyberattack

High Reliability Organizations

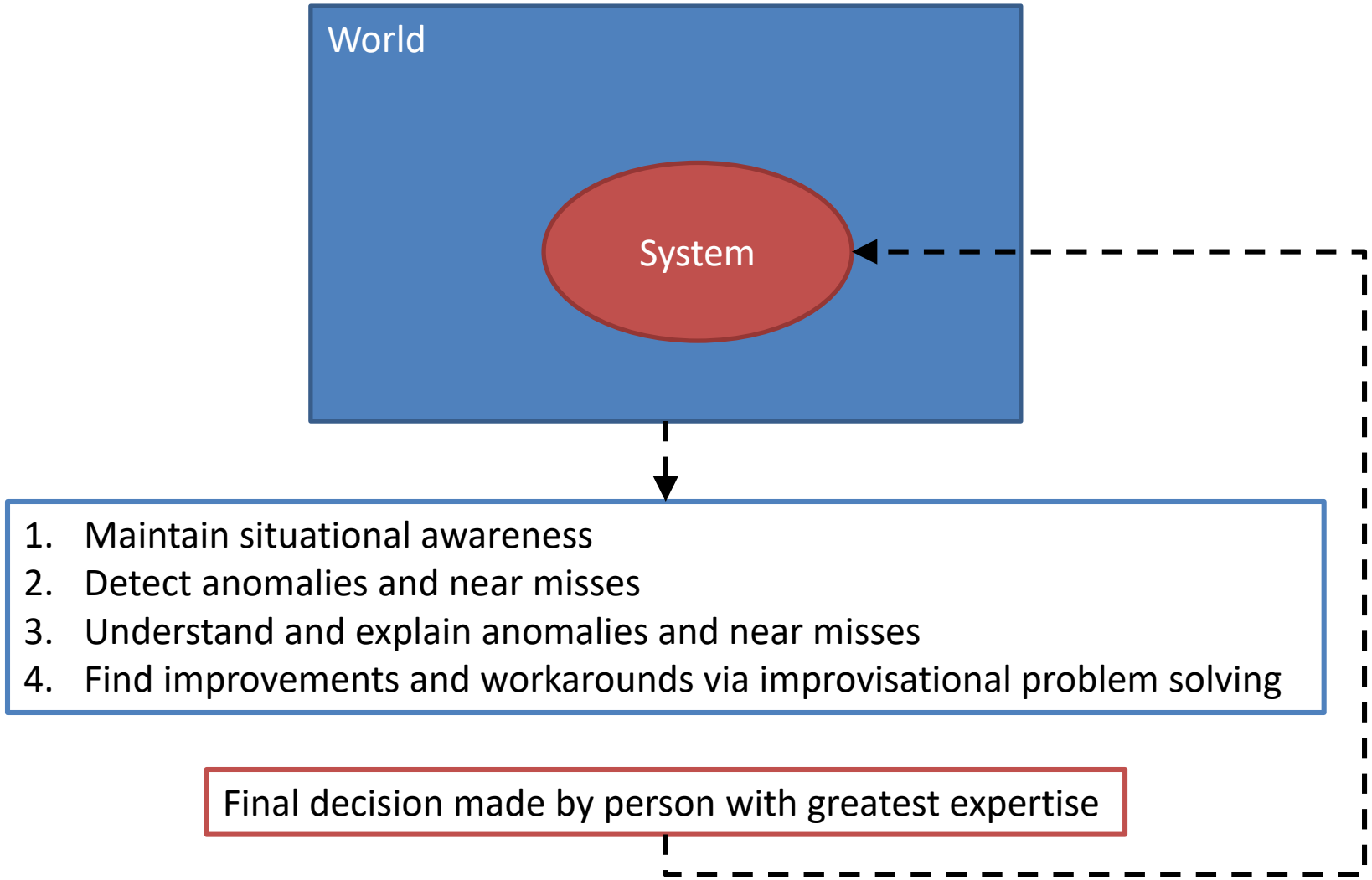
Todd LaPorte, Gene Rochlin, and Karlene Roberts

- Studied several high reliability human teams
  - Air Traffic Control
  - Nuclear power plant operations
  - Aircraft Carrier Flight Deck Operations
- Claim: Accidents can be prevented through organizational design, culture, management, and human choices
- Impact:
  - Patient safety movement
  - Cockpit resource management

Properties of High Reliability Organizations

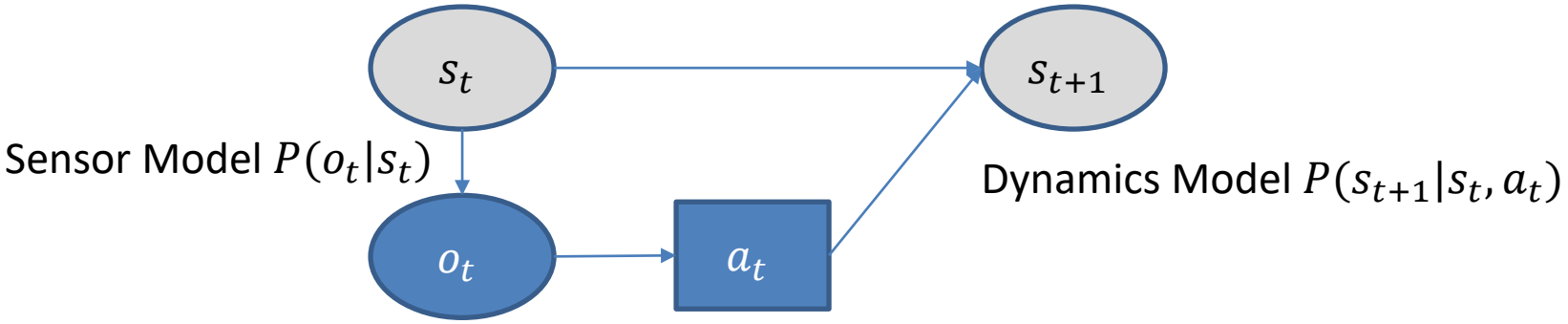
- Preoccupation with failure
  - Fundamental belief that the system has unobserved failure modes
  - Treat anomalies and near misses as symptoms of a problem with the system
- Reluctance to simplify interpretations
  - Comprehensively understand the situation
- Sensitivity to operations
  - Maintain continuous situational awareness
- Commitment to resilience
  - Develop the capability to detect, contain, and recover from errors. Practice improvisational problem solving
- Deference to expertise
  - During a crisis, authority migrates to the person who can solve the problem, regardless of their rank

Principle: There are unknown failure modes



PART 1: AUTONOMOUS AI SYSTEMS

Maintain Situational Awareness



- Maintain a probability distribution $P(s_t)$ over the state of the system
- Collect the observations o_t
- Compute updated distribution:

$$P(s_t|o_t) \propto P(o_t|s_t)P(s_t)$$

- Choose the action a_t
- Predict next state distribution:

$$P(s_{t+1}|o_t, a_t) = \sum_{s_t} P(s_{t+1}|a_t, s_t)P(s_t|o_t)$$

- Methods:
  - Kalman filter
  - Particle filters
  - Expectation propagation
  - Variational approximations
  - etc.

Detect Anomalies and Near Misses

Detecting Anomalies

- Compute the "surprise" of the observed o_{t+1}
- Predicted distribution of o_{t+1}:

$$P(o_{t+1}|o_t, a_t) = \sum_{s_{t+1}} P(s_{t+1}|o_t, a_t) P(o_{t+1}|s_{t+1})$$

- Anomaly Score:

$$-\log P(o_{t+1}|o_t, a_t)$$

- Practical algorithms may require approximations

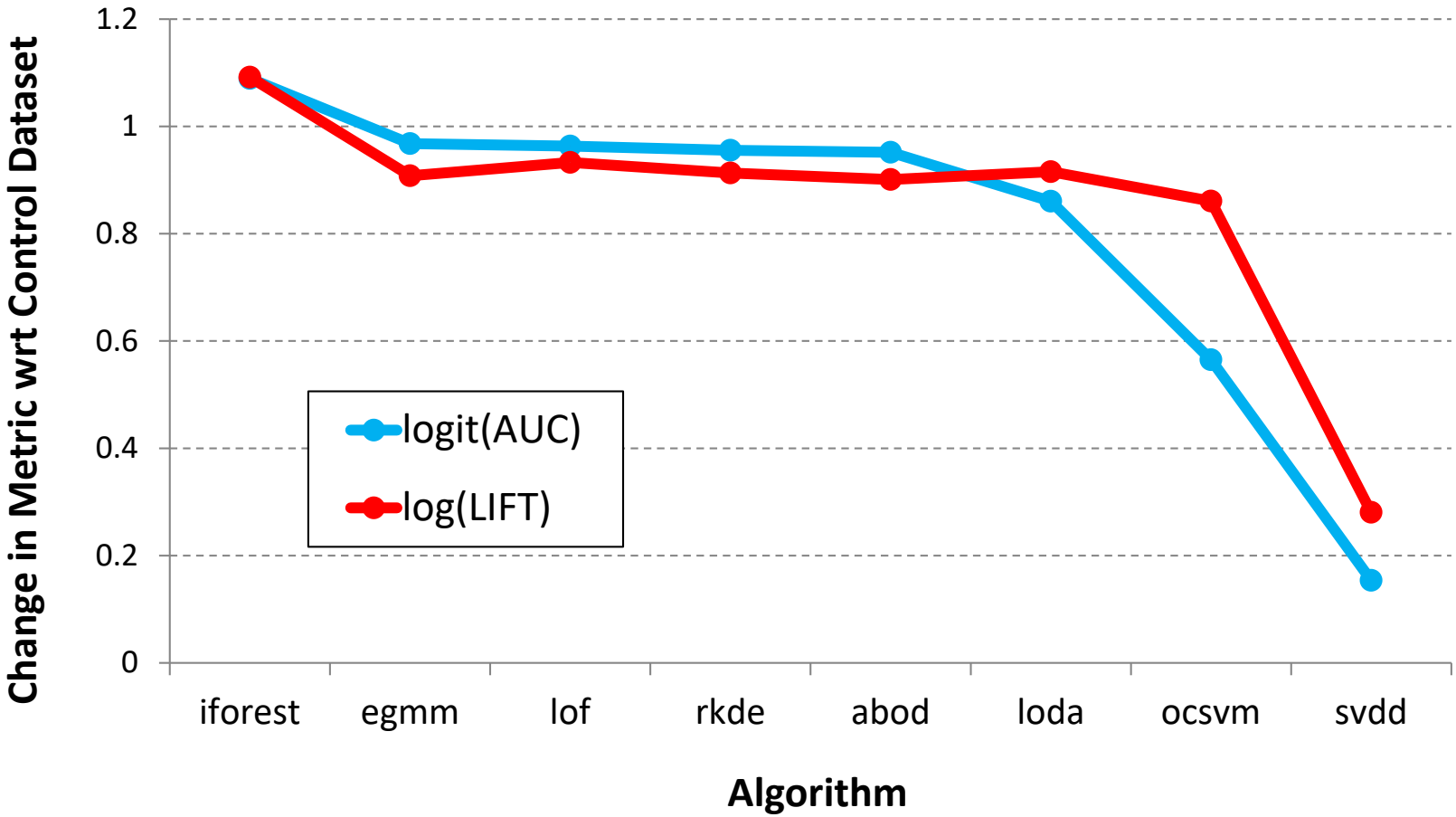
Anomaly Detection Benchmarking Study

- Goal: Compare published algorithms on a robust collection of benchmarks
  - Previous comparisons suffered from small size and/or proprietary data sets

- **Density-Based Approaches**
  - RKDE: Robust Kernel Density Estimation (Kim & Scott, 2008)
  - EGMM: Ensemble Gaussian Mixture Model (our group)
- **Quantile-Based Methods**
  - OCSVM: One-class SVM (Schoelkopf, et al., 1999)
  - SVDD: Support Vector Data Description (Tax & Duin, 2004)
- **Neighbor-Based Methods**
  - LOF: Local Outlier Factor (Breunig, et al., 2000)
  - ABOD: kNN Angle-Based Outlier Detector (Kriegel, et al., 2008)
- **Projection-Based Methods**
  - IFOR: Isolation Forest (Liu, et al., 2008)
  - LODA: Lightweight Online Detector of Anomalies (Pevny, 2016)

[Emmott, Das, Dietterich, Fern, Wong, 2013; KDD ODD-2013]
[Emmott, Das, Dietterich, Fern, Wong. 2016; arXiv 1503.01158v2]

Anomaly Detection Benchmark Results

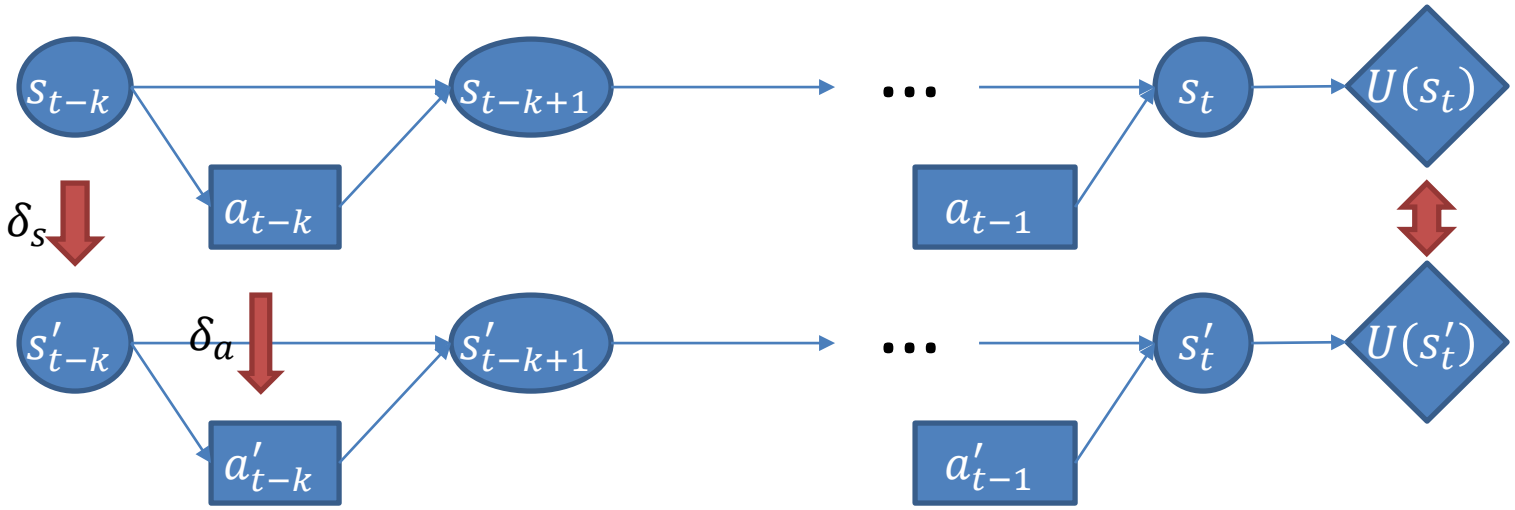


iForest was best; quantile methods were worst; all others approximately equal

Anomaly Detection Challenges

- High-dimensional spaces are inherently difficult
  - Can we assume the true state s has much lower dimension?
- Image and video data
  - Need to discover the lower-dimensional space
  - Discover the dynamics model $P(s_{t+1}|s_t, a_t)$
  - Discover the sensor model $P(o_t|s_t)$
- Promising directions
  - Auto-encoders and generative models (VAE, RAE, BiGAN)
  - Neural Rendering Model
  - Extending existing methods to work with time series

Defining and Detecting Near Misses



- Suppose we have a utility function $U(s)$ over states
- Counterfactual Notion: Perturb s_{t-k} and/or a_{t-k}
- Near Miss:

$$U(s'_t) \ll U(s_t)$$

- Detecting near misses is under-studied; requires causal model
- Should anticipate them and act to prevent them (ACAS-X)

Explaining Anomalies and Near Misses: Research Challenges

- Open-ended space of hypotheses
  - Effects of exogenous variables / unknown external agents?
    - what external agents might exist and why would they be affecting our system?
  - Sensor failures and/or inadequate sensors
    - why didn't we detect the anomaly or near miss earlier?
  - Model failures (dynamics and sensor models)
    - did the system structure change? (broken pipe? stuck valve?)
- Promising work
  - Model-based diagnosis including performing information-gathering actions

Finding Repairs and Workarounds

- Approaches
  - Update dynamics and sensor models and then apply planning algorithms?
  - Mark aspects of the models as unreliable and seek a plan that does not depend on those aspects?
  - Always plan conservatively to be robust to model errors?
- Existing Work
  - Optimizing Against an Adversary (robust optimization)
    - Robust Optimization
      - Ben-Tal, Bertsimas, etc.
  - Optimizing Conditional Value at Risk (CVaR)
    - Acting conservatively provides robustness to model error

Summary: Autonomous AI

	Assessment
Situational Awareness	A mature methods
Detect Anomalies and Near Misses	B high-dimension, dynamics
Explain Anomalies and Near Misses	D only basic techniques
Improvise Solutions	F

PART 2: AI + HUMAN TEAMS

AI and Human Teams

- Even very powerful AI systems will be surrounded by a human team that will determine
  - What goals to give it
  - What degree of autonomy to permit it
  - When to trust it
  - What degree of learning/adaptation to allow

- How can the combined AI + Human Team be safe and robust?
  - Reconsider each aspect of high-reliability organizations from an interactive perspective

Situational Awareness: Past Failures

- Autopilot Tunnel Vision: Aircraft autopilot not aware of air traffic control instructions
  - Co-pilot must continually update the autopilot's waypoints based on ATC interactions
  - This load increases in high-traffic/high-risk situations
  - Co-pilot loses awareness of other aspects of the system
- Autopilot Fails to Communicate Situation
  - Colgan Air 3407 crash near Buffalo
  - Autopilot was compensating for aircraft icing, but pilots were not aware of this
  - Eventually autopilot was forced to hand control back to pilots
  - Their lack of situational awareness led to crash ("decompensation failure")
- Autopilot Over-Communicates
  - Hundreds of unimportant alarms
  - Complex displays that bury important information
- Humans Misunderstand Internal State of Autonomous System
  - USS John McCain collision: team thought single slider was controlling both engines, but it was controlling only one
  - Caused ship to turn into the course of an oncoming ship

Requirements for Robust Situational Awareness

- AI system should have sufficient sensing
  - state of world including other agents
  - state of the system being controlled
  - state of its human team
- Human team and AI system should establish and maintain a shared mental model
  - AI system should reason about what the users know and do not know and communicate strategically
  - Humans need a good mental model of the AI system's beliefs about the situation
  - AI system needs to be able to explain its beliefs to humans
  - Careful design of user interface is critical

Anomaly and Near Miss Detection

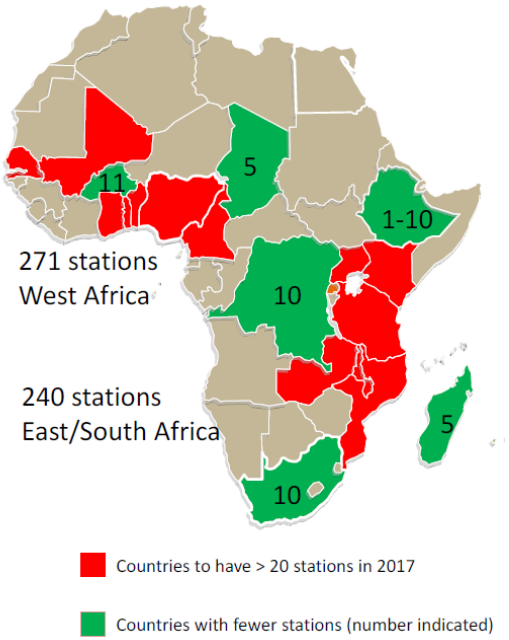
- Existing methods are highly local
  - sensor readings out of standard range
  - violations of minimum separation (air-to-air, air-to-ground, car-to-car)
- Need more and better anticipation of problems
  - model the behavior of other agents (including team members)
  - project system state many steps into the future and evaluate
- Incorporate interactive anomaly detection

Explaining Anomalies and Near Misses

- Existing anomaly explanations are purely statistical
  - "This credit card transaction is anomalous because it was very large compared to this customer's normal behavior"
- Root cause analysis
  - "Customer just purchased a house and is buying furniture for it"
  - Must consider a broader set of hypotheses than in normal state updating
  - May lack dynamics and observation models for this broader space

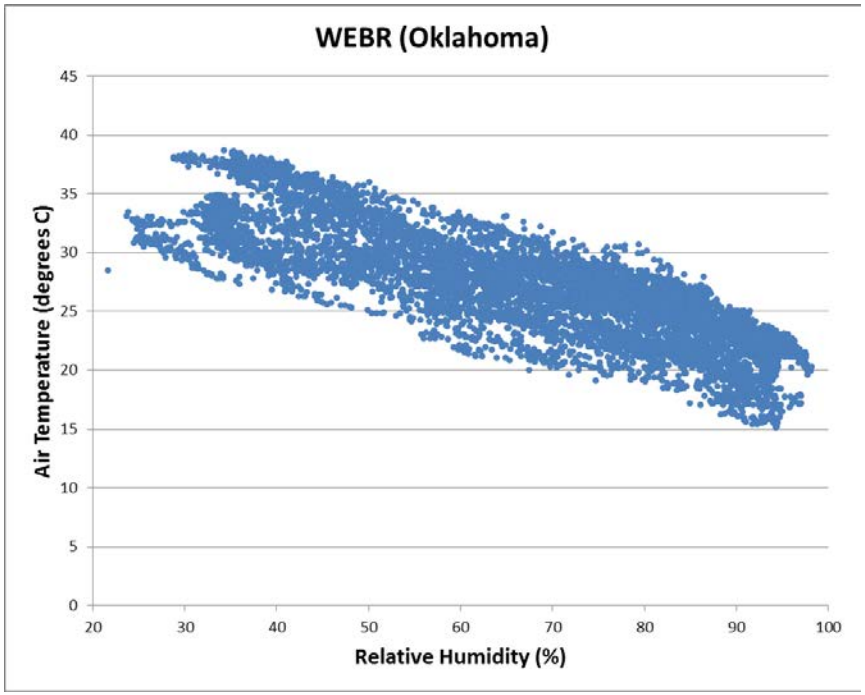
Example: Root Cause Analysis in TAHMO

- TAHMO: Trans-Africa Hydro-Meteorological Observatory
  - 500+ automated weather stations in East and West Africa
  - Data quality control: Detect broken sensors

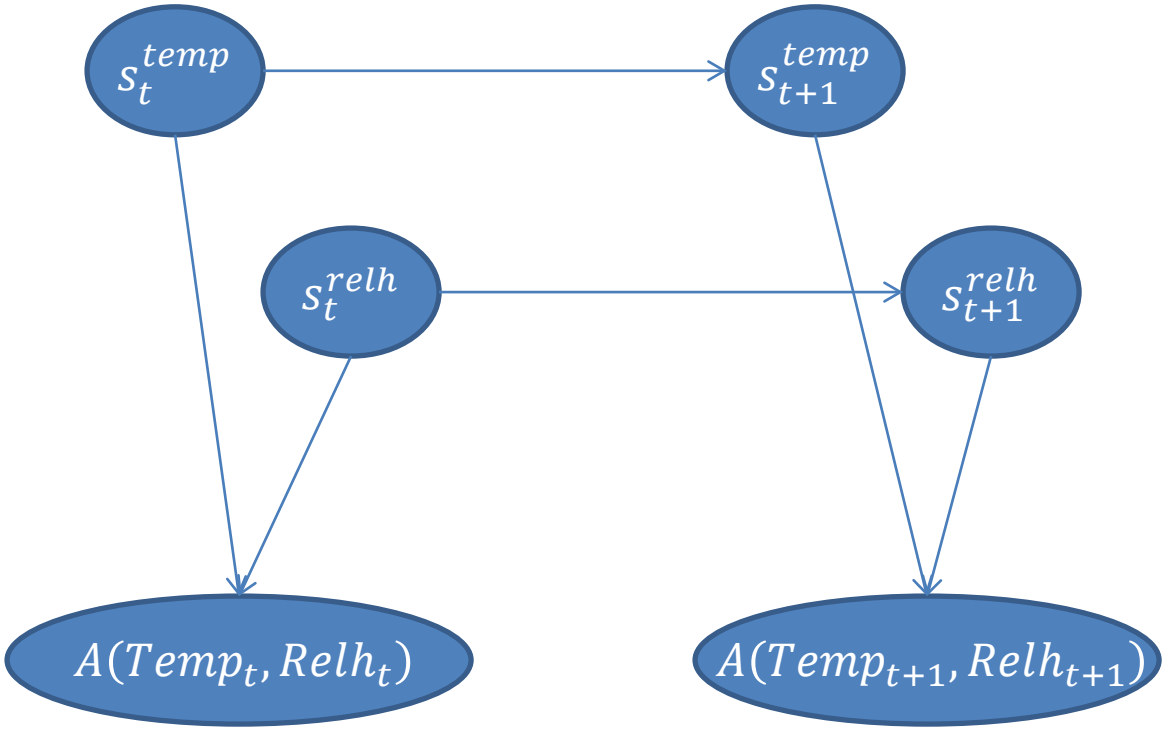


Detect Anomalies from Violated Correlations

- Joint distribution of temperature and relative humidity $P(T, RH)$
- Anomaly has high $-\log P(T, RH)$
- But how do we know which sensor (thermometer vs. humidity) is broken?
- Solution: probabilistic inference over multiple views



Joint Anomaly Detection

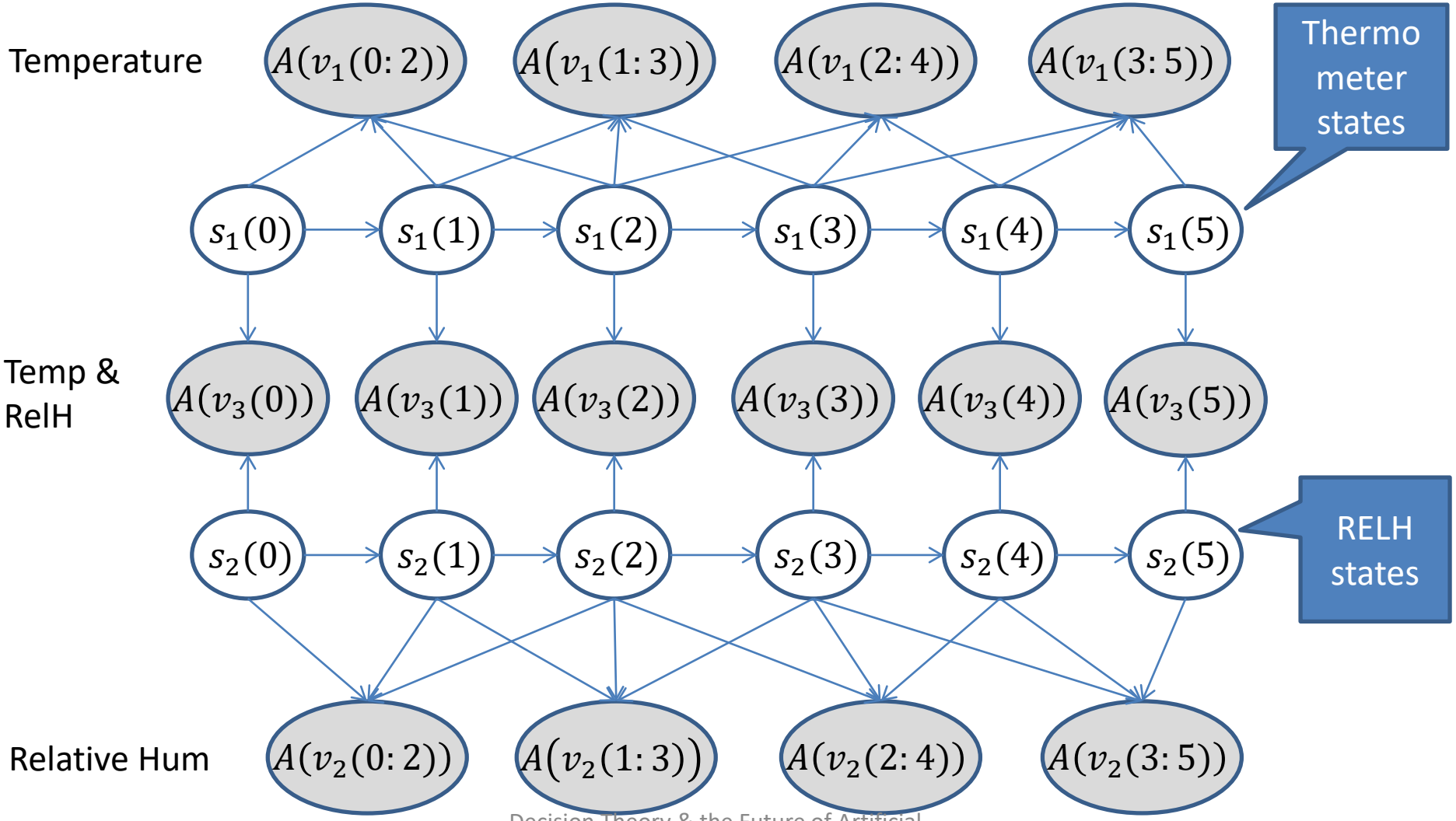


SENSOR-DX: Multiple View Approach

- Views capture joint distribution over time and space
  - Single sensor over K time steps
    - $A(x_{t-K+1}, x_{t-K+1}, \ldots, x_{t-1}, x_t)$ captures this distribution
  - Pair of sensors at one time step
    - $A(x_t, y_t)$ such as temperature and relative humidity
  - Difference between value at station ℓ and the value predicted from spatial neighbors $\ell'_1, \ldots, \ell'_k$
    - $A\left(x_t(\ell) - f\left(x_t(\ell'_1), \ldots, x_t(\ell'_k)\right)\right)$

Diagnostic Model

Which sensor states best explain the observed anomaly scores?



Probabilistic Root Cause Analysis

- Assemble incoming data into view tuples
- Compute anomaly score for each view tuple
- Perform probabilistic inference to determine which sensor states best explain the observed anomaly scores:

$$\arg\max_{S} P(S|A(V))$$

- Challenge: How to explain the inferred root causes?
  - "If (temperature, relative humidity) combination is anomalous but temperature time series looks nominal, then the relative humidity sensor must be broken"

Improvisational Problem Solving

- Human users and AI system collaborate to find solutions
- Humans “think outside the box” to enlarge the problem space
- How can the AI system help humans reason about this larger problem space?
  - Verify that proposed plan does not violate any known system limits or lead to bad system states within the AI’s narrow problem space?
  - Can humans communicate the larger space to the AI system so that it can reason about it?
  - Explain to humans how the AI system would behave if permitted autonomy
- Existing work:
  - Mixed-initiative Planning

Mixed-Initiative Planning

- Agenda of activities that need to be planned
- User-invoked planning operators
  - Plan all: fully automated planning
  - Plan selected goals: incrementally add one or more activities to the emerging plan
  - Expand selected subgoal
  - Create a plan sketch (commit to some activities, possibly at different levels of abstraction)
- User plan editing
  - Move an activity to a different time while disturbing existing steps as little as possible
  - Add/delete activity
  - Delete or relax a constraint
  - Tentatively fix a decision but note that if additional information arrives (e.g., weather forecast) then this decision should be revisited
- System continually checks that all constraints are satisfied and makes changes to satisfy resource constraints and mutual exclusion constraints

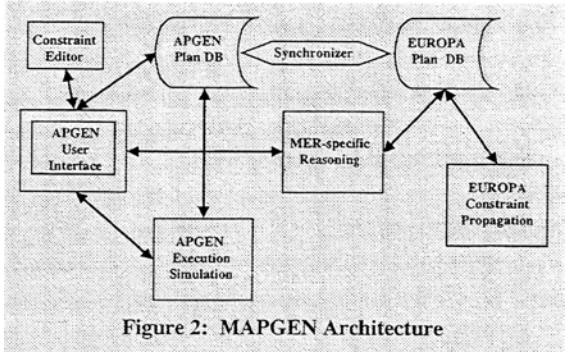
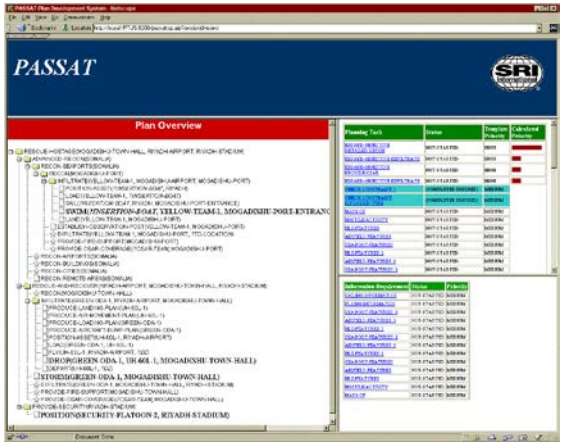


Figure 2: MAPGEN Architecture

MAPGEN: Bresina et al., 2005



PASSAT: Myers, et al., 2002

Decision Making

- Person with the relevant expertise should make the final decision
  - Course of action
  - Decision to delegate actions to AI system

Past Failures

- AI capabilities and limitations are unclear to humans
  - Humans trust AI autonomy when they should not
    - Gulf War Patriot Missile Fratricide
      - New crew operating unfamiliar equipment
      - Broken radio communication with other teams
      - Patriot missile system incorrectly interpreted returning British fighter jet as incoming ballistic missile
      - Crew trusted the system, launched defensive missile: 2 killed
    - Iran-Iraq War AEGIS autonomous ship defense system
      - AEGIS and crew misinterpreted civilian aircraft as incoming attacker despite IFF transponder signal
      - Armed AEGIS which then shot down the aircraft: 290 killed

- AI current and future behavior is difficult for humans to predict
  - Symptom: Humans continually monitor AI system behavior and prepare to intervene at any moment

Past Failures (2)

- Teamwork failures lead to accidents
  - USS Fitzgerald collision with ACX Crystal
    - Poor communications including failure to use advanced navigation aides led to loss of situational awareness
    - Collision killed 7

Requirements for Human + AI Teams

- AI System needs to monitor functioning of human team
  - Detect communication failures
  - Detect misunderstandings (failures of shared mental model)
- AI System needs to know when to defer to human expertise
  - Model the expertise of each team member
  - Know whom to engage to obtain information or make a decision
- If human teamwork is breaking down, AI system should abort mission and switch to a safe backup plan

Summary: Human + AI Teams

	Assessment
Situational Awareness	C poor UI, poor communication
Detect Anomalies and Near Misses	C user feedback to anomaly detection
Explain Anomalies and Near Misses	D only basic techniques
Improvise Solutions	D mixed-initiative planning

CONCLUSIONS

Pure AI Research Needs

- Learning and modeling of complex, partially-sensed systems
  - latent dynamical models
- Anomaly detection for high-dimensional and latent dynamical systems
- Near miss prediction and detection
- Root cause diagnosis of near misses and anomalies
- Robust planning for incompletely-understood systems

Human + AI Research Needs

- Joint situational awareness – Shared Mental Models
  - Learning and reasoning about what other agents (human and AI) know and believe
- Interactive anomaly detection
- Interactive near miss detection
- Joint improvisational planning
- Joint expertise model
  - AI has model of expertise of each team member
  - Team members can accurately predict the behavior of the AI system and known when to trust it (and when not)
- Explanation and visualization methods to support all of the above

Granting Autonomy is a Form of Trust

- Being trustworthy is more than being predictable and reliable
- Trust carries with it several obligations
  - To act on behalf of the team's goals and interests
  - To keep the team well-informed
  - To return control to the team when it cannot meet these obligations

QUESTIONS?

“Normal Accidents”

Charles Perrow (1984)

- Response to Three-Mile Island failures
- Claims:
  - Accidents are inevitable (“normal”) in extremely complex systems
  - If system also has catastrophic potential, these accidents will lead to catastrophe

Impact: Patient Safety Movement

- Goal: Zero Preventable Deaths in Health Care
- Checklists in the operating room
- Empowering all members of the surgical team to halt the surgery if a problem is noticed

Culture of Safety	Healthcare-associated Infections (HAIs)	Medication Safety
Monitoring for Respiratory Depression	Patient Blood Management	Hand-off Communications
Neonatal Safety	Airway Safety	Early detection and treatment of Sepsis
Prevention/Resuscitation of Cardiac Arrest	Obstetric Safety	Embolic Events
Mental Health	Fall Prevention	Nasogastric Tube Placement & Verification
Person & Family Engagement	Patient Safety Curriculum	Post-operative Delerium in older adults

HRO Desideratum for AI Deployment

We should not deploy AI unless we can ensure that the human organization is highly reliable