Toward High-Reliability Artificial Intelligence

Thomas G. Dietterich, Distinguished Professor (Emeritus)

Former President, AAAI (Association for the Advancement of Artificial Intelligence)

Former President, IMLS (International Machine Learning Society)

Fellow AAAI, ACM, AAAS

Oregon State University, Corvallis, OR USA 97331

Oregon State University

High-Reliability Human Organizations (HROs)

Todd LaPorte, Gene Rochlin, and Karlene Roberts (Weick, et al., 1999)

Organizations that achieve consistently low error rates over long periods of time

Prior belief: There are unknown failure modes in the system. The role of the organization is to discover and fix them

1. Maintain continuous situational awareness
2. Detect anomalies and near misses
3. Generate and evaluate multiple hypotheses
4. Design, implement, and test solutions
5. Final decision is made by the person with the most expertise

Impact of HRO Principles

- Cockpit Resource Management:
  - Train aircraft pilots and co-pilots to detect and recover from novel failures
- Patient Safety Movement:
  - Eliminate all preventable medical mistakes
- My Goal: AI Safety Movement
  - Eliminate all preventable AI mistakes

The Problem

Medical Errors

Claim the Lives of 3+ Million Patients Every Year

Globally it is believed that medical errors kill more people than HIV, Malaria, and Tuberculosis, combined. COVID-19 exposed all the deficiencies in healthcare that already existed and placed patients and health workers at greater risk for preventable harm.

The Vision

ZERO Preventable Harm and Death

ZERO is not just a number – it's our mission

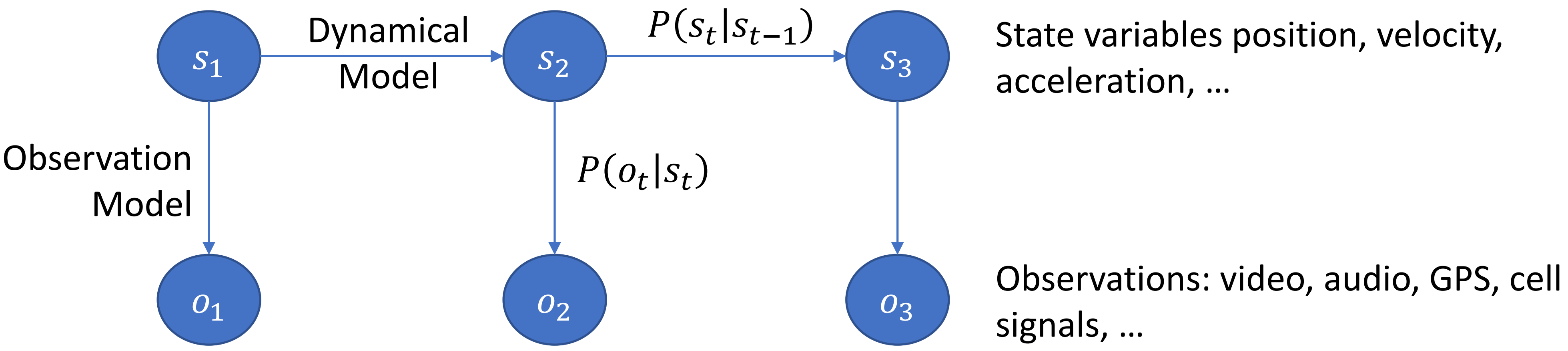
The Patient Safety Movement Foundation believes reaching ZERO preventable patient harm and deaths across the globe by 2030 is not only the right goal, but an attainable one with the right people, ideas, and technology.

http://patientsafetymovement.org

Lessons for AI: Two Scenarios

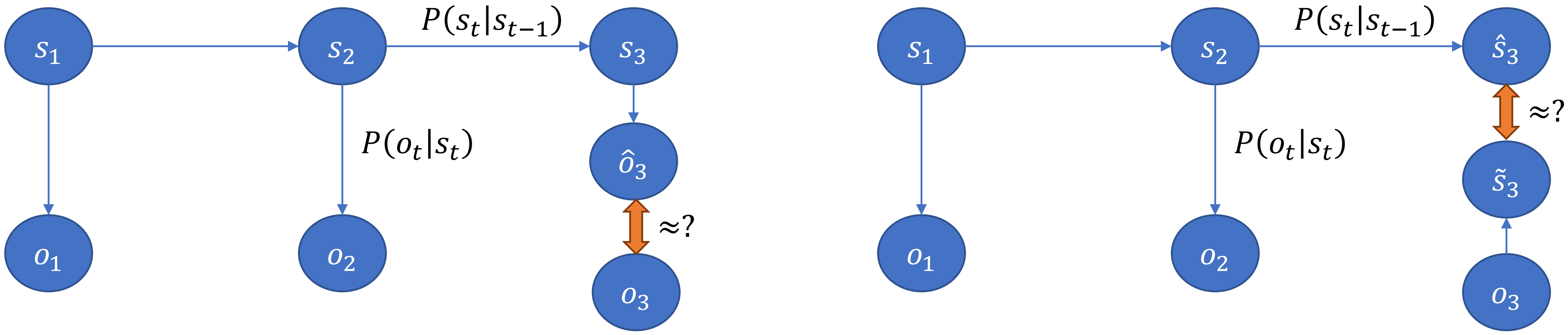
- Scenario 1: Autonomous AI System as an HRO
- Scenario 2: Human organization + AI System as an HRO

Designing AI Systems to be HROs: Situational Awareness



- Probabilistic models provide a powerful and well-understood method for tracking the state of a system
- $P(s_t|o_1, \dots, o_t) = P(s_0)\prod_{u=1}^{t} P(s_u|s_{u-1})P(o_u|s_u)$

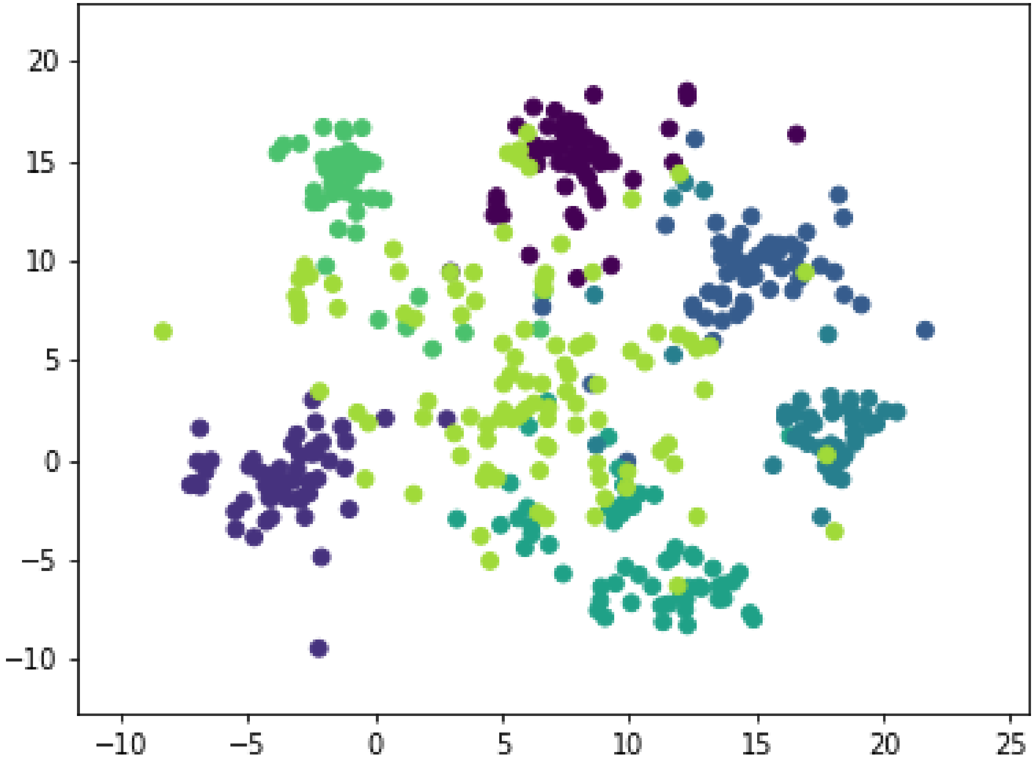
Detecting Anomalies as Violated Expectations



- Compare predicted observation $P(\hat{o}_t|s_t)$ to actual observation o_t
  - Surprisal: $-\log P(o_t|s_t)$
  - Distance: $||\hat{o}_t - o_t||$
- Difference in state space: compare $P(\hat{s}_t|s_{t-1})$ to $P(\tilde{s}_t|o_t)$
  - Point difference: $||\hat{s}_t - \tilde{s}_t||$
  - Distributional distance: $TV\big(P(\hat{s}_t|s_{t-1}), P(\tilde{s}_t|o_t)\big)$

Deep Learning Challenge: Learned Representations

- Deep Learning creates its own representations (e.g., for s and o)
- These do not always give useful distances or probability densities
- This makes it difficult to compute
  - $||s - \hat{s}||$
  - $||o - \hat{o}||$
  - etc.



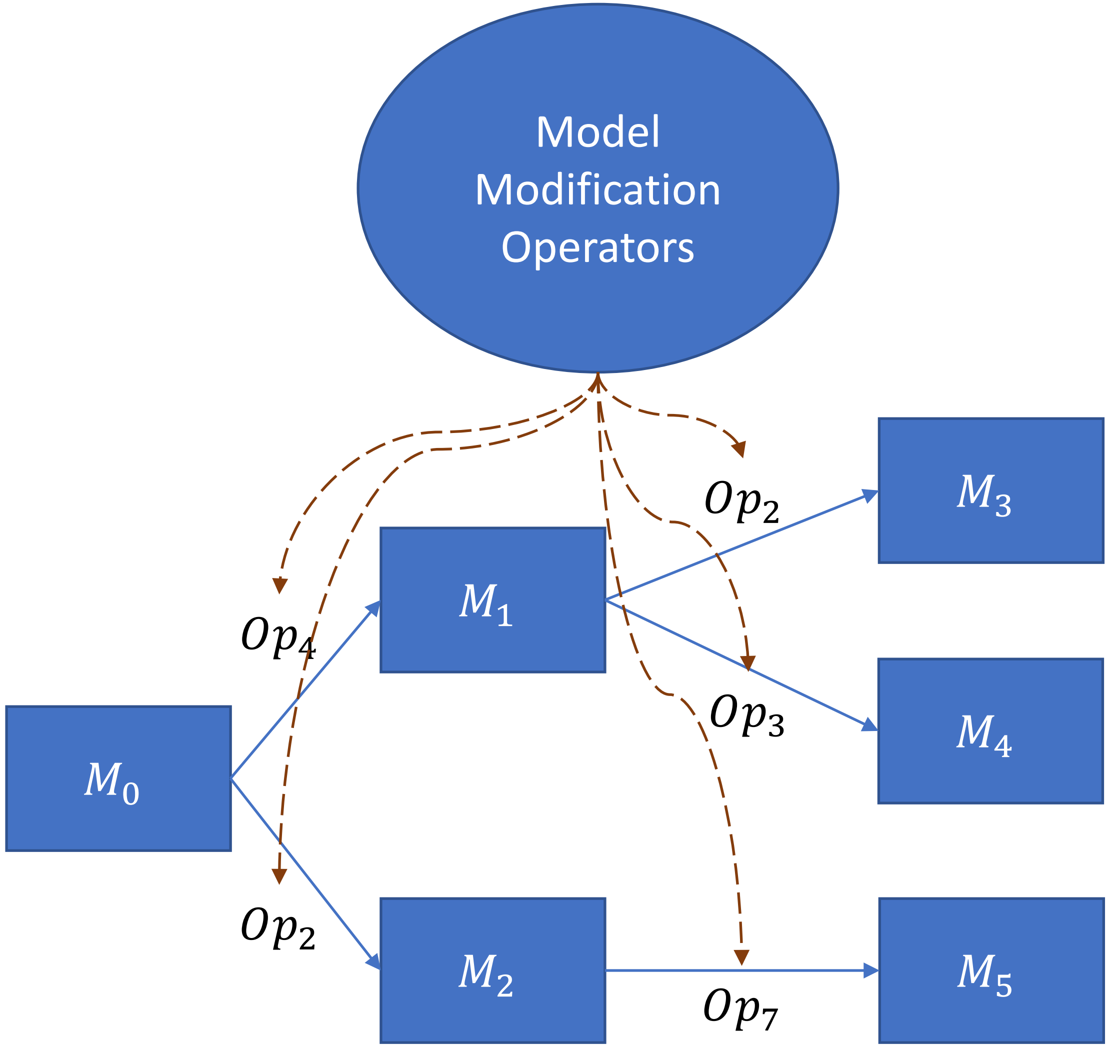
Learned representations of
- 6 known categories (dark colors)
- 4 novel categories (light green)

Detecting Near Misses

- Near Miss: “If I had taken a slightly different action, the result would have been much worse”
  - This is a “counter-factual” statement and requires a causal model (Pearl, 2009)
  - The model can be analyzed to measure sensitivity of the outcome to small changes in the inputs
  - Very little research in this area

Formulating Hypotheses and Implementing Fixes

- Search in a space of model changes to find a model that can account for the observations
- What if no such modification can be found?
  - Can AI systems “think outside the box”?
  - Hypothesize and refine novel properties, structures, relationships, causal pathways?
- Existing work focuses on a narrow set of models. Example: Digital circuits
  - Substitute gate (NAND →AND)
  - Add/remove connection
  - Add/remove path to ground



Designing a Human + AI Team to be an HRO

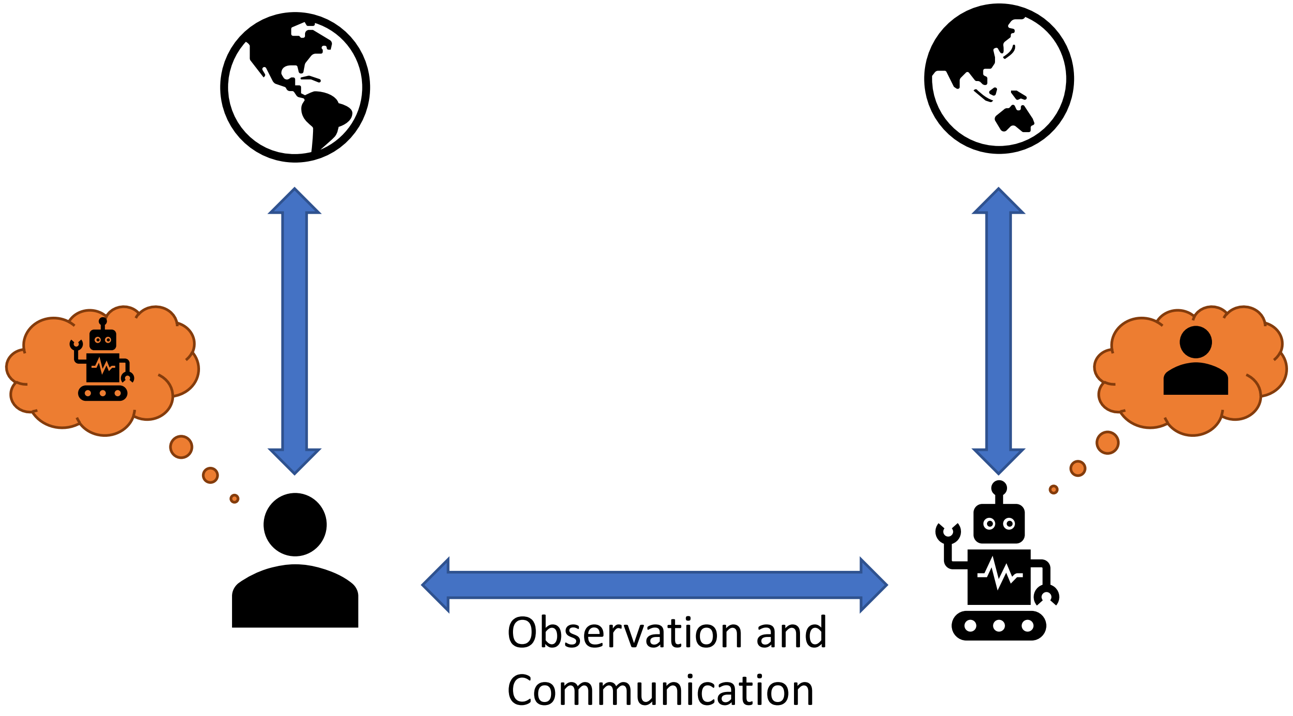
- Every AI system will be surrounded by a human team
- Risk that inserting AI technology into a human HRO may destroy its reliability
- Historical example: Aircraft Autopilots
  - Autopilot is a primitive type of AI system
  - The autopilot had poor situational awareness
    - co-pilot needed to enter waypoint coordinates into the autopilot system
  - Pilots had poor insight into the state of the autopilot system
  - Result: crashes during hand-off from autopilot to human pilots

Colgan Air Flight 3407

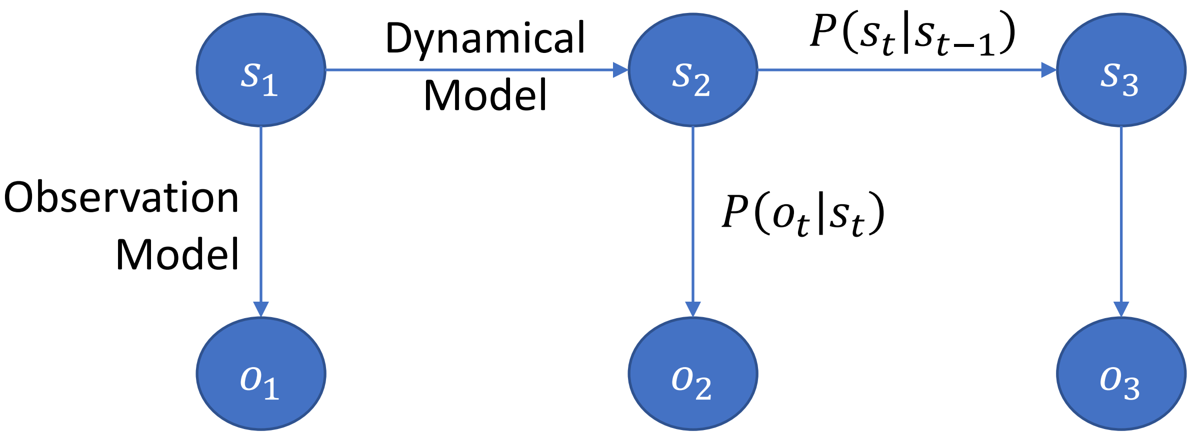
https://www.thedickinsonpress.com/news/1780177-investigator-plane-fell-flat-buffalo-house

Human-AI Situational Awareness

- The Human is aware of some aspects of the world
- The AI is aware of some aspects of the world
- Human must have an accurate model of the AI
- AI must have an accurate model of the human



Anomaly and Near Miss Detection



State variables include state of each human and AI

Observations: video, audio, chat between team members

- AI must model the humans as well as the physical system
  - detect unusual behavior of the humans
  - detect near miss events (e.g., where humans almost made a serious mistake)

Formulating Hypotheses and Implementing Fixes

- Human and AI system must work together to construct and evaluate hypotheses
- How can the humans communicate hypotheses to the AI system?
  - The AI system needs to be able to *represent* outside-the-box hypotheses
- How can the AI system communicate its hypotheses to the humans?
  - Hypotheses must be interpretable to the humans
- How can the AI system support improvisational problem solving?
  - Probably need to practice working with humans

Missing Technology for Human-AI HROs

- Deep language understanding
- Complex conversations
- Recursive Theory of Mind
- Methods for expanding the model space
- Models of human improvisational problem solving
- Models of human behavior (especially in high-stress situations)

Fundamental Conclusion

- **AI should not be deployed (either autonomously or as part of a human-AI joint team) until we can assure that they are highly reliable**
- We must develop standards and measurement techniques for evaluating the reliability of AI and joint Human-AI systems
- In certain applications, it may not be feasible to achieve high-reliability
  - self-driving cars?
  - autonomous weapons systems?
  - autonomous medical systems?
- We should not deploy AI in those applications
  - risk of catastrophic failures

Bibliography

- Dietterich, T. G. (2019). Robust artificial intelligence and robust human organizations. *Frontiers in Computer Science*, *13*(1), 1–3.
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why.* Basic Books.
- Weick, K., Sutcliffe, K., & Obstfeld, D. (1999). Organizing for high reliability: Processes of collective mindfulness. In R. S. Sutton & B. M. Staw (Eds.), *Research in Organizational Behavior* (Vol. 1, pp. 81–123). Jai Press.