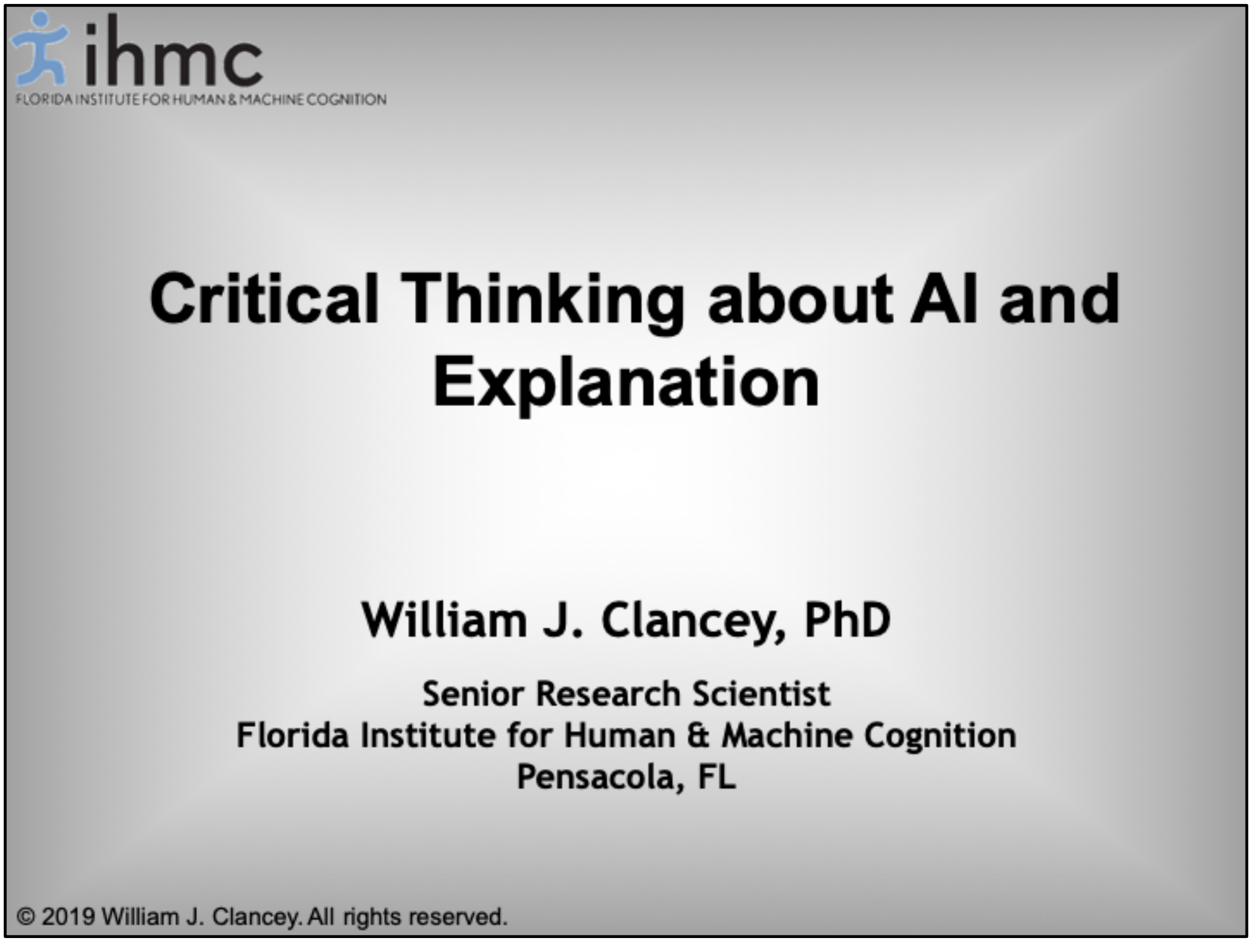


ihmc

FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION

Critical Thinking about AI and Explanation

William J. Clancey, PhD

Senior Research Scientist
Florida Institute for Human & Machine Cognition
Pensacola, FL

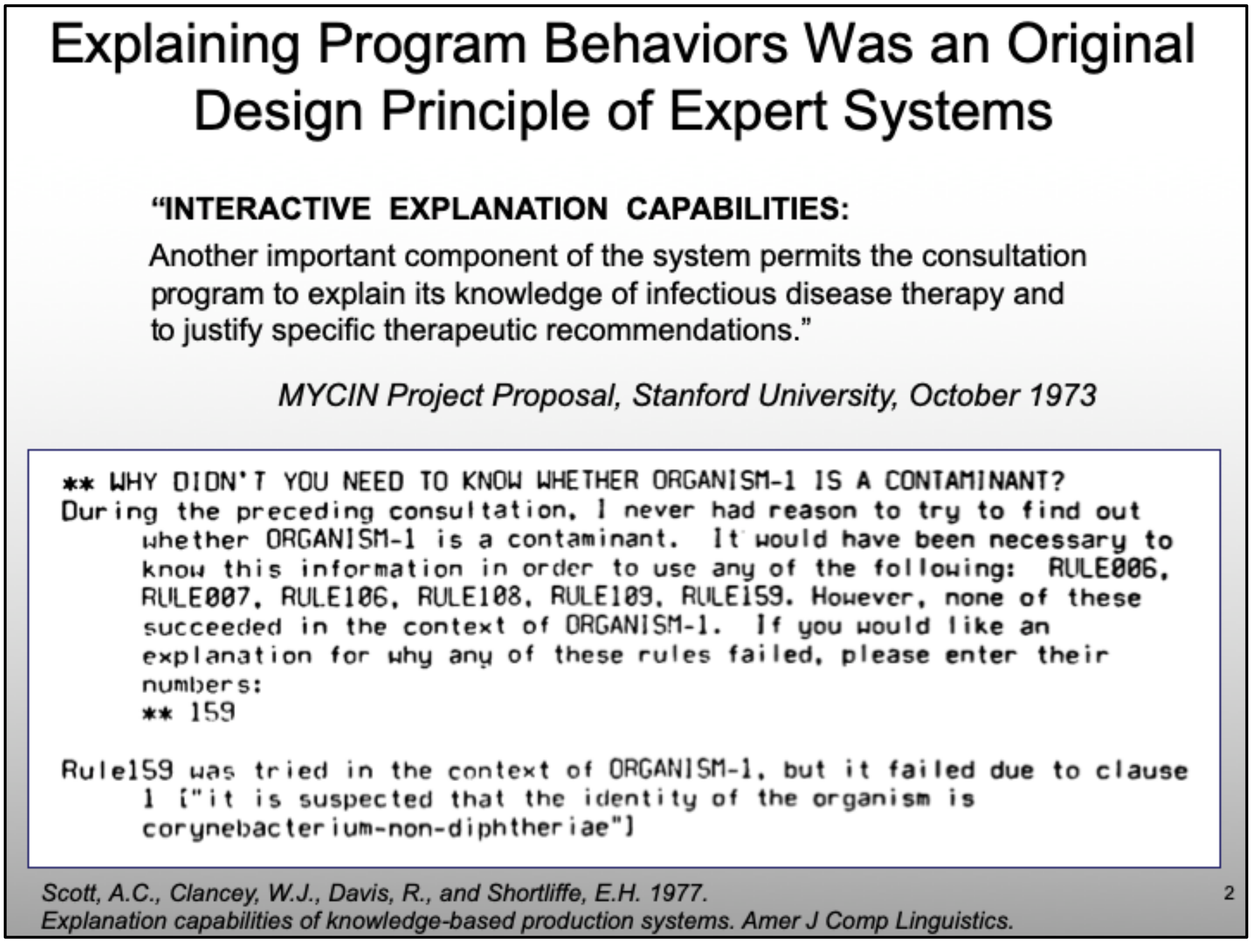
This panel’s concern with explanation is important, but we need to view the topic broadly.

I’d like to make two points.

First, AI programming encompasses far more than what’s called “machine learning” today. Automated systems will typically include computational models of processes, what we called “symbolic AI” in the 1970s and 80s.

Second, the problems posed by advanced automation requires much more than providing explanations, the entire R&D process needs to be revamped.

When I say “critical thinking,” I mean locating AI systems squarely in the language and methods of science and engineering.



Explaining Program Behaviors Was an Original Design Principle of Expert Systems

> **"INTERACTIVE EXPLANATION CAPABILITIES:**
> Another important component of the system permits the consultation program to explain its knowledge of infectious disease therapy and to justify specific therapeutic recommendations."

MYCIN Project Proposal, Stanford University, October 1973

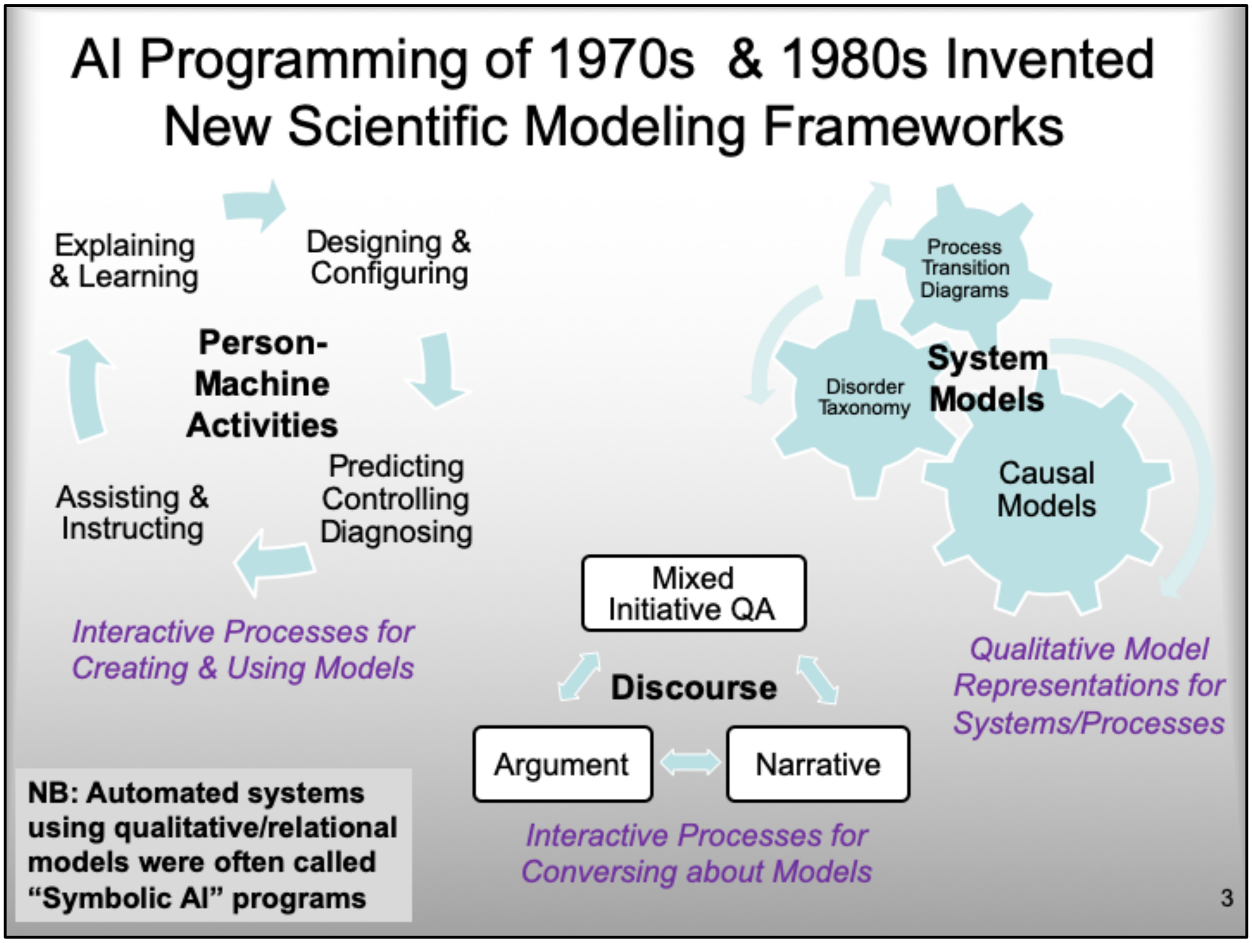
```
** WHY DIDN'T YOU NEED TO KNOW WHETHER ORGANISM-1 IS A CONTAMINANT?
During the preceding consultation, I never had reason to try to find out
     whether ORGANISM-1 is a contaminant.  It would have been necessary to
     know this information in order to use any of the following:  RULE006,
     RULE007, RULE106, RULE108, RULE109, RULE159. However, none of these
     succeeded in the context of ORGANISM-1.  If you would like an
     explanation for why any of these rules failed, please enter their
     numbers:
     ** 159

Rule159 was tried in the context of ORGANISM-1, but it failed due to clause
     1 ["it is suspected that the identity of the organism is
     corynebacterium-non-diphtheriae"]
```

Scott, A.C., Clancey, W.J., Davis, R., and Shortliffe, E.H. 1977.
Explanation capabilities of knowledge-based production systems. Amer J Comp Linguistics.

In expert systems projects of the 1970s, explanation capability was treated as an obvious design requirement.

For example, MYCIN explained decisions in terms of data and inferences, -- how it made a conclusion, why it requested patient data, why it didn't conclude or ask about something….and so on.

(This was actually my first publication.)



AI Programming of 1970s & 1980s Invented New Scientific Modeling Frameworks

Explaining & Learning

Designing & Configuring

Person-Machine Activities

Assisting & Instructing

Predicting Controlling Diagnosing

Interactive Processes for Creating & Using Models

Process Transition Diagrams

Disorder Taxonomy

System Models

Causal Models

Qualitative Model Representations for Systems/Processes

Mixed Initiative QA

Discourse

Argument

Narrative

Interactive Processes for Conversing about Models

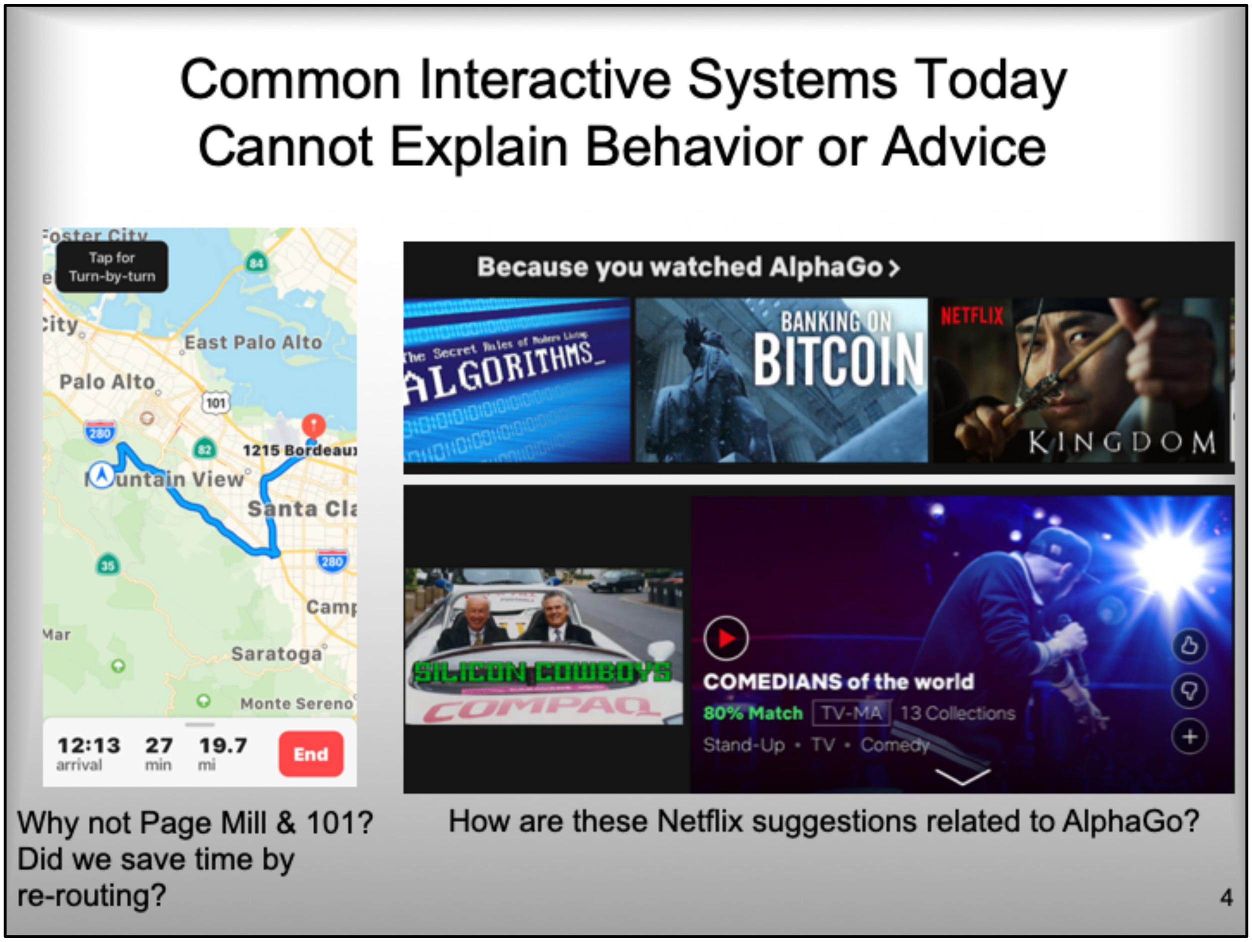
NB: Automated systems using qualitative/relational models were often called "Symbolic AI" programs

3

The most important thing to know about expert systems and symbolic AI in general is that the programs used computer-interpretable models of processes.

In our attempt to replicate human intelligence we made a major contribution to science and engineering by developing computational modeling methods that represent processes and systems in relational languages — typical examples are semantic networks, conceptual classifications, and causal networks.

I am concerned that when people refer to the ”failure” of symbolic AI they do not understand the nature of these system modeling methods and their ongoing contribution. These methods are not made obsolete by neural network programs anymore than statistics replaces our need for causal theories in science, engineering, and medicine.



Common Interactive Systems Today Cannot Explain Behavior or Advice

Why not Page Mill & 101? Did we save time by re-routing?

How are these Netflix suggestions related to AlphaGo?

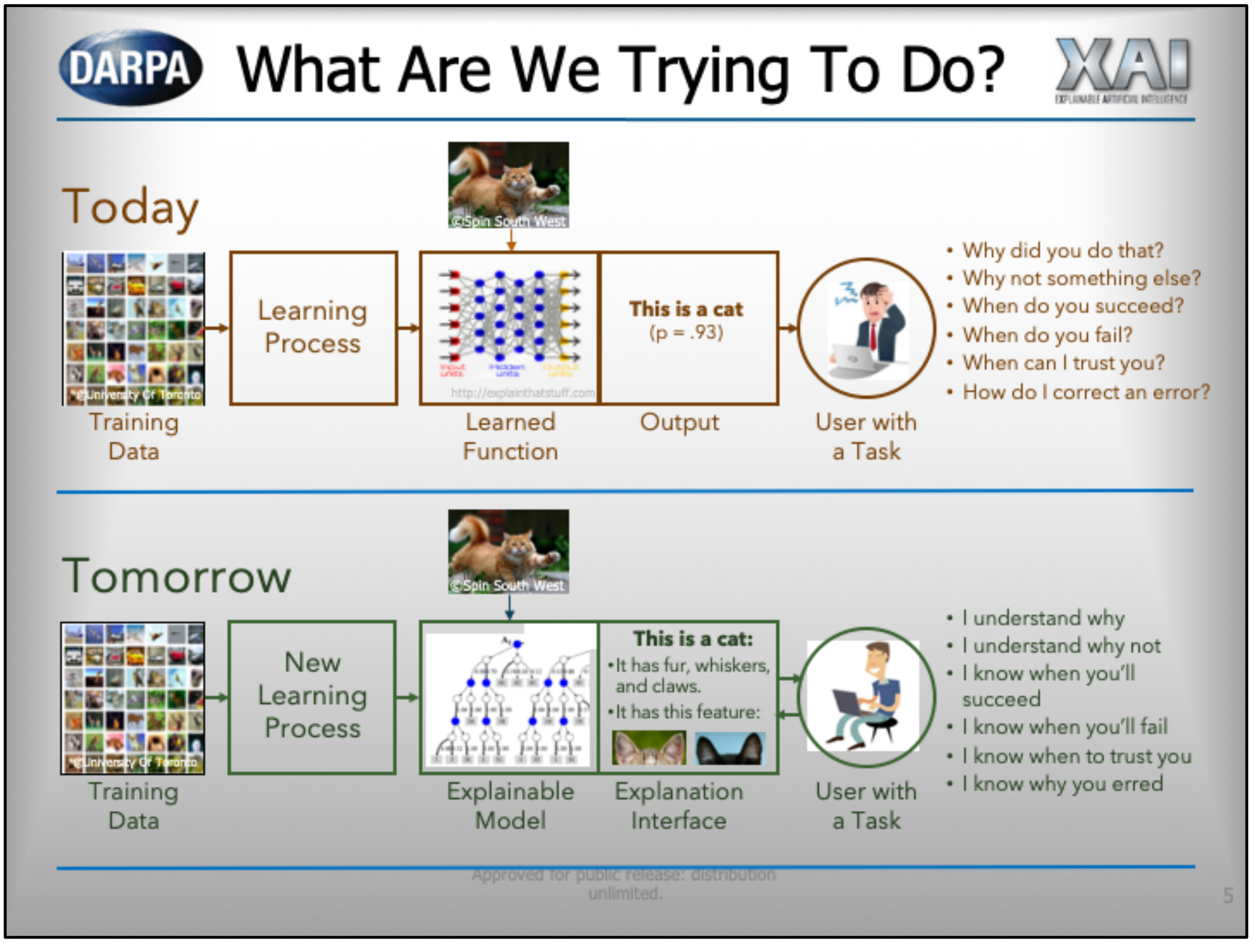
4

So where are we after 40 years of AI explanation research? Here are two examples of AI programs used by millions of people everyday.

Why can't I ask the iPhone Maps program simple questions about where we are going? After it recommends changing the route, why can't it review how much time we saved, so I know whether to trust it next time?

Netflix's advice program is also a black box. What do these movie suggestions have to do with my watching AlphaGo? (Netflix could filter and sort results by inferring how movies are related.)

The lack of explanation is a real problem for the programmers, too, who when they verify and improve these programs. We used MYCIN"s explanation system every day in refining our disease and diagnostic models. Apple's and Netflix's programs are primitive compared to state-of-the-art expert systems in the 1970s. Consumers might not care, but DARPA has recognized the problem.



What Are We Trying To Do?

Today

Training Data → Learning Process → Learned Function → Output: **This is a cat** (p = .93) → User with a Task

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow

Training Data → New Learning Process → Explainable Model → Explanation Interface: **This is a cat:** •It has fur, whiskers, and claws. •It has this feature: → User with a Task

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

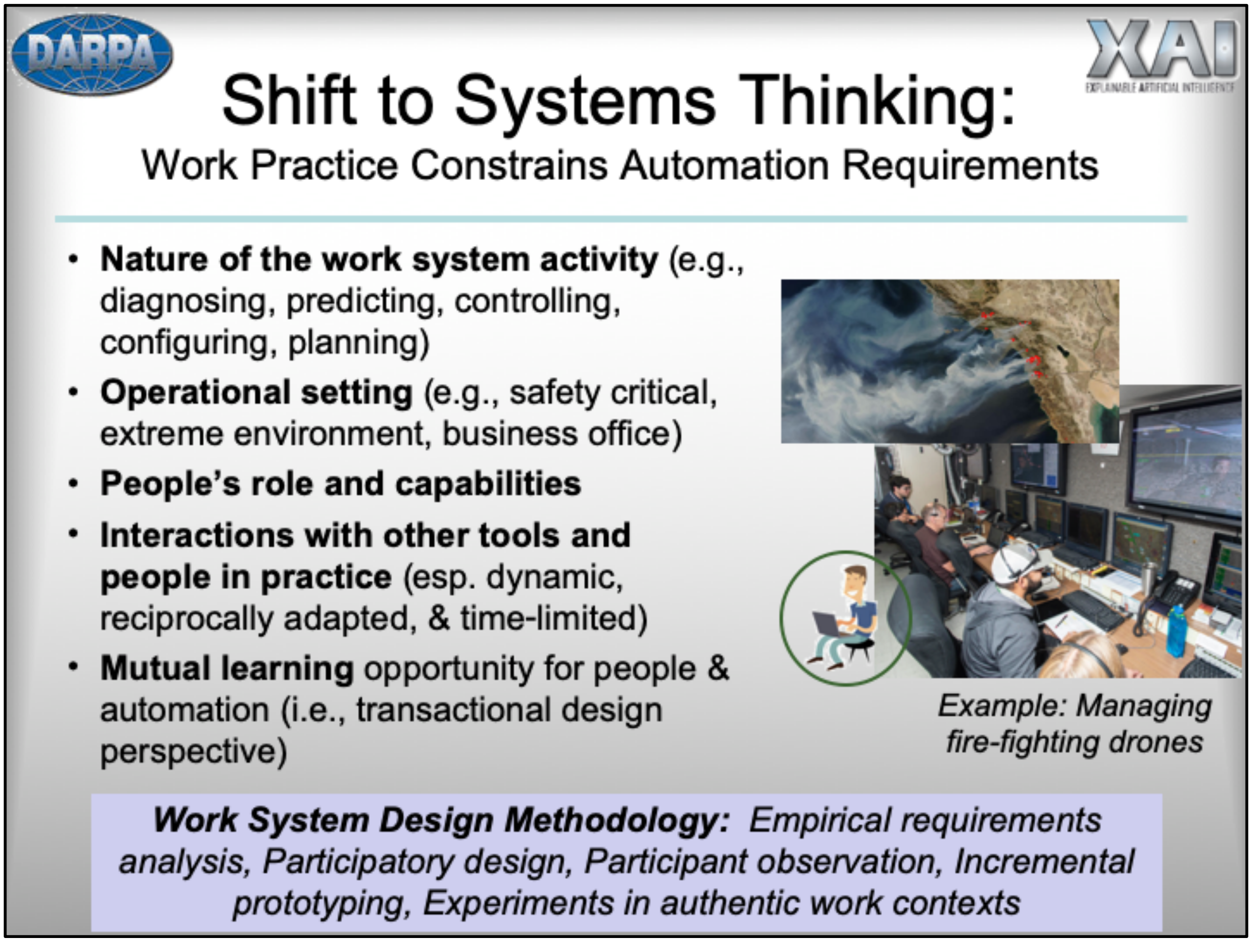
Approved for public release: distribution unlimited.

5

Slide from 2018-2019 presentations of David Gunning, DARPA XAI Program Manager

As you heard earlier, the "Explainable AI" (XAI) research program is creating new modeling methods that can generate explanations. But are they understandable and useful?

What constitutes "an explainable model" depends on the context of the person's activity, which is not considered in this initial exploratory phase of XAI research.



DARPA
XAI
EXPLAINABLE ARTIFICIAL INTELLIGENCE

Shift to Systems Thinking:

Work Practice Constrains Automation Requirements

- **Nature of the work system activity** (e.g., diagnosing, predicting, controlling, configuring, planning)
- **Operational setting** (e.g., safety critical, extreme environment, business office)
- **People's role and capabilities**
- **Interactions with other tools and people in practice** (esp. dynamic, reciprocally adapted, & time-limited)
- **Mutual learning** opportunity for people & automation (i.e., transactional design perspective)

Example: Managing fire-fighting drones

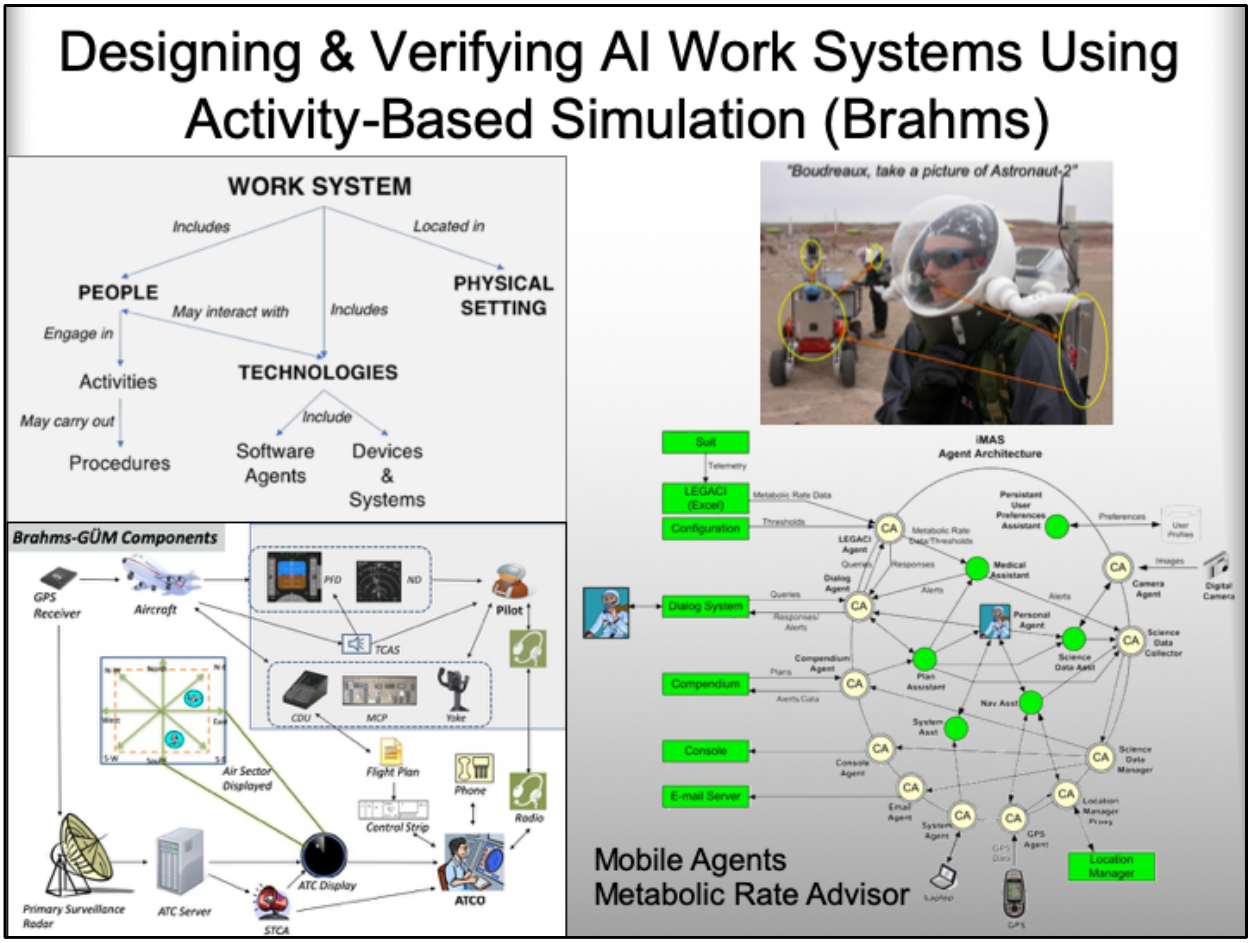
***Work System Design Methodology:** Empirical requirements analysis, Participatory design, Participant observation, Incremental prototyping, Experiments in authentic work contexts*

The next phase of XAI research will need be reoriented—"explanation" is not just a module—it must be an integral part of the work system's design.

Explanation and all interactions with people are constrained by the nature of the work, the setting, other people's roles, and so on.

Is this a real-time control activity in which a decision must be made in the next few seconds or minutes? Or is this a long-term planning activity that allows days or weeks for interacting with the program?

Methods for developing sociotechnical systems have matured. We have learned that what people need to understand and what might be automated will be discovered in experiments with system prototypes. At NASA I called this methodology "Empirical Requirements Analysis." We have also learned that an especially effective design method is modeling the entire work system in a computer simulation.



Designing & Verifying AI Work Systems Using Activity-Based Simulation (Brahms)

Brahms-GÜM Components

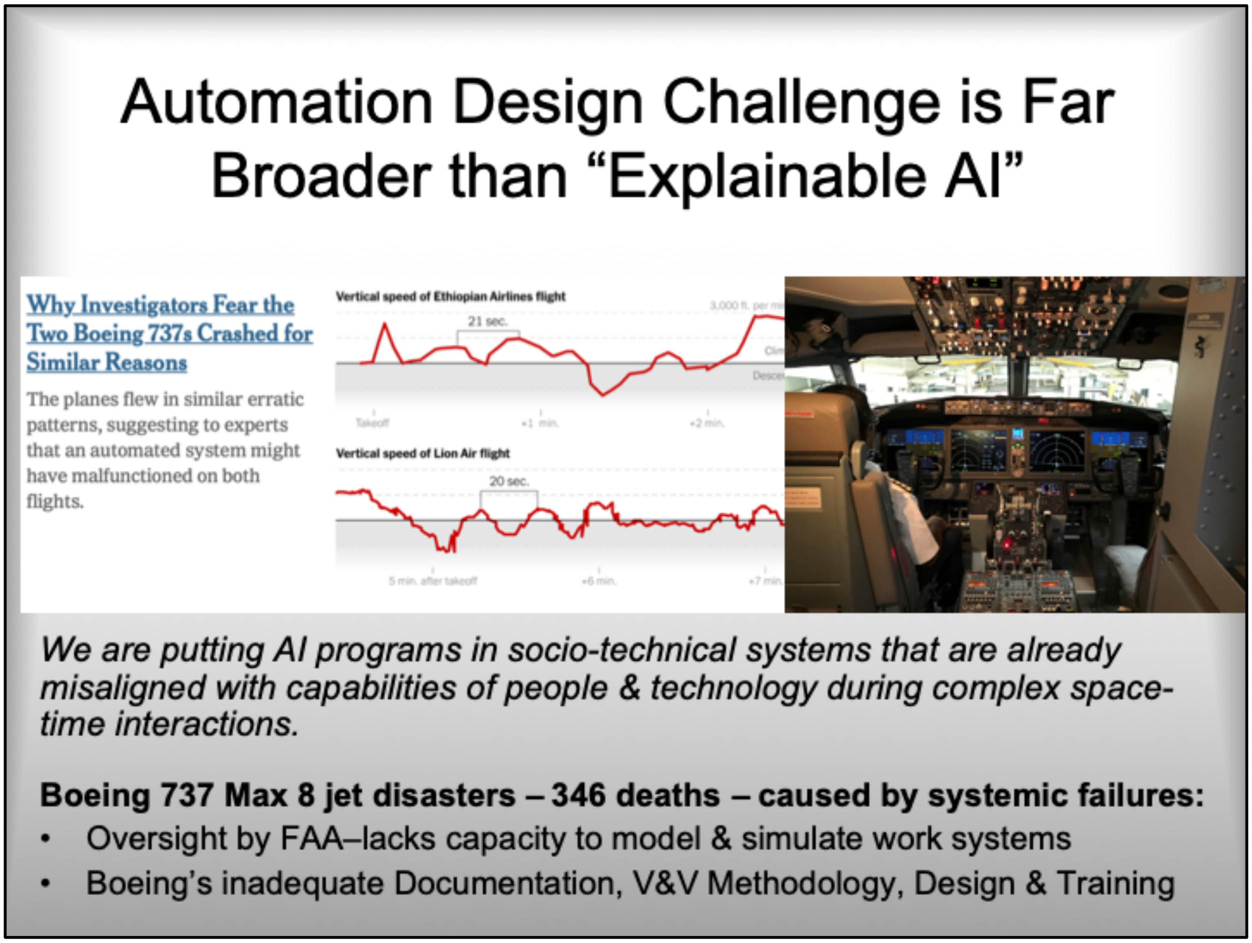
"Boudreaux, take a picture of Astronaut-2"

Mobile Agents
Metabolic Rate Advisor

At NASA we used the Brahms work practice simulation framework– in which all processes are modeled independently and interact in a simulated environment. A key idea is modeling people's behaviors, their activities, not abstracted functions or tasks.

For example, in Brahms-GÜM we simulated how the Traffic Collision Avoidance System (TCAS) interacts with pilots as they are interacting with Air Traffic Controllers to understand emergent time-space interactions. This was part of NASA Ames research on the nature of authority and trust in mixed systems of people and AI agents.

We developed a series of voice-commanded agent systems in the Mobile Agents Project, which demonstrated how to design, implement, and refine agents using a comprehensive work system simulation. This design approach, focusing on people's practices, is broadly applicable to developing automated systems.



Automation Design Challenge is Far Broader than "Explainable AI"

Why Investigators Fear the Two Boeing 737s Crashed for Similar Reasons

The planes flew in similar erratic patterns, suggesting to experts that an automated system might have malfunctioned on both flights.

Vertical speed of Ethiopian Airlines flight
3,000 ft. per mi
21 sec.
Clim
Desce
Takeoff
+1 min.
+2 min.
Vertical speed of Lion Air flight
20 sec.
5 min. after takeoff
+6 min.
+7 min.

We are putting AI programs in socio-technical systems that are already misaligned with capabilities of people & technology during complex space-time interactions.

Boeing 737 Max 8 jet disasters – 346 deaths – caused by systemic failures:

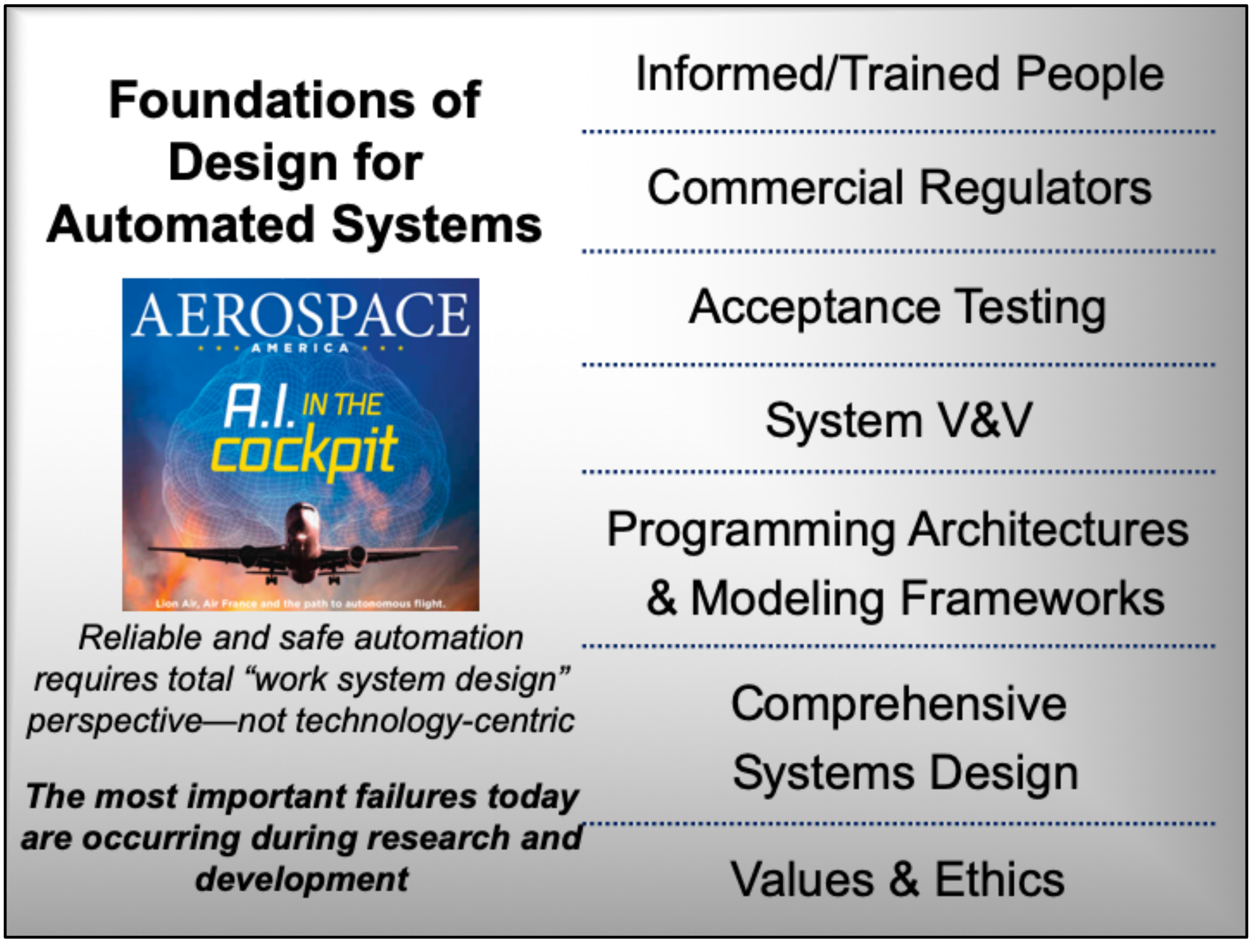
- Oversight by FAA–lacks capacity to model & simulate work systems
- Boeing's inadequate Documentation, V&V Methodology, Design & Training

We see in the news everyday that the challenge posed by advanced automation is far broader than today's XAI research considers.

Adding an explanation module to Boeing 737 Max 8 would not have prevented the planes from crashing.

It is becoming clear that the design, certification, and training methods used by vendors and the FAA are inadequate for today's automated systems.

In particular, the Max 8 illustrates how adding new forms of automation to an existing system may cause complex interactions with people to emerge in practice.


AEROSPACE
AMERICA
A.I. IN THE
cockpit
Lion Air, Air France and the path to autonomous flight.

Foundations of Design for Automated Systems

Reliable and safe automation requires total "work system design" perspective—not technology-centric

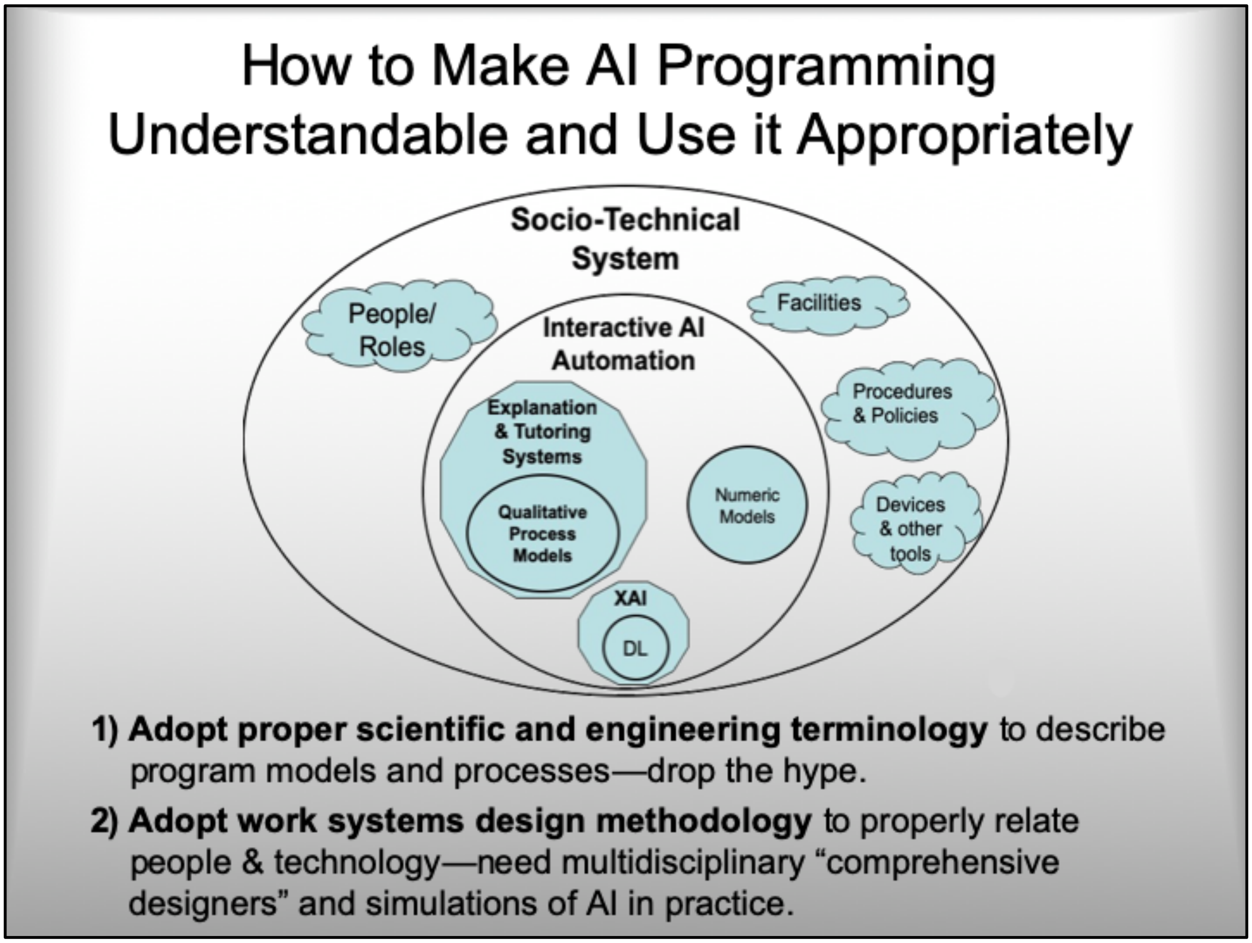
The most important failures today are occurring during research and development

- Informed/Trained People
- Commercial Regulators
- Acceptance Testing
- System V&V
- Programming Architectures & Modeling Frameworks
- Comprehensive Systems Design
- Values & Ethics

Human factors has generally focused on failures that occur during operations. But I would claim that the most important failures today are occurring during R&D.

We must breakout of the technology-centric perspective. We need more tools like Brahms that simulate not just the machinery and programs, but include how people interact with automated systems in practice.

This includes any explanation capability that is intended to be part of the work system.



How to Make AI Programming Understandable and Use it Appropriately

Socio-Technical System

People/ Roles

Facilities

Procedures & Policies

Devices & other tools

Interactive AI Automation

Explanation & Tutoring Systems

Qualitative Process Models

Numeric Models

XAI

DL

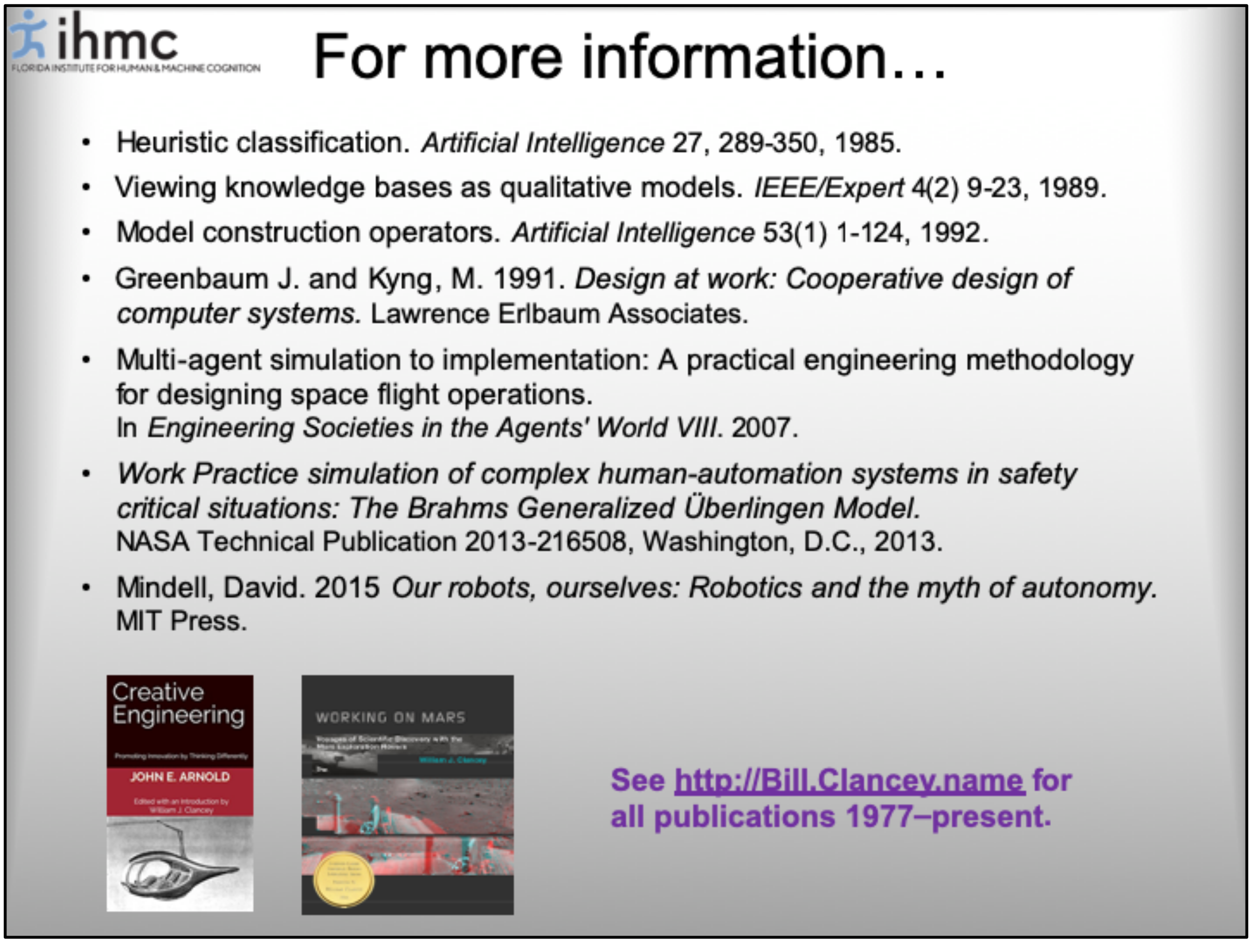
1) **Adopt proper scientific and engineering terminology** to describe program models and processes—drop the hype.

2) **Adopt work systems design methodology** to properly relate people & technology—need multidisciplinary “comprehensive designers” and simulations of AI in practice.

So to sum up, my two points are that what’s called Deep Learning or Neural networks today will usually be just a small part of the automation in a work system, and providing explanation is just a part of the problem of developing automation that fits how people think and work.

We need a comprehensive work system design approach and tools for designing and certifying advanced automation systems.

And our problem is not just creating explainable systems, but better characterizing and explaining the entire AI enterprise. We should adopt a proper scientific and engineering terminology. For example, don’t speak about ”neural networks” unless you are modeling the brain.



ihmc
FLORIDA INSTITUTE FOR HUMAN & MACHINE COGNITION
Creative Engineering
JOHN E. ARNOLD
WORKING ON MARS

For more information…

- Heuristic classification. *Artificial Intelligence* 27, 289-350, 1985.
- Viewing knowledge bases as qualitative models. *IEEE/Expert* 4(2) 9-23, 1989.
- Model construction operators. *Artificial Intelligence* 53(1) 1-124, 1992.
- Greenbaum J. and Kyng, M. 1991. *Design at work: Cooperative design of computer systems.* Lawrence Erlbaum Associates.
- Multi-agent simulation to implementation: A practical engineering methodology for designing space flight operations.
  In *Engineering Societies in the Agents' World VIII*. 2007.
- *Work Practice simulation of complex human-automation systems in safety critical situations: The Brahms Generalized Überlingen Model.*
  NASA Technical Publication 2013-216508, Washington, D.C., 2013.
- Mindell, David. 2015 *Our robots, ourselves: Robotics and the myth of autonomy.* MIT Press.

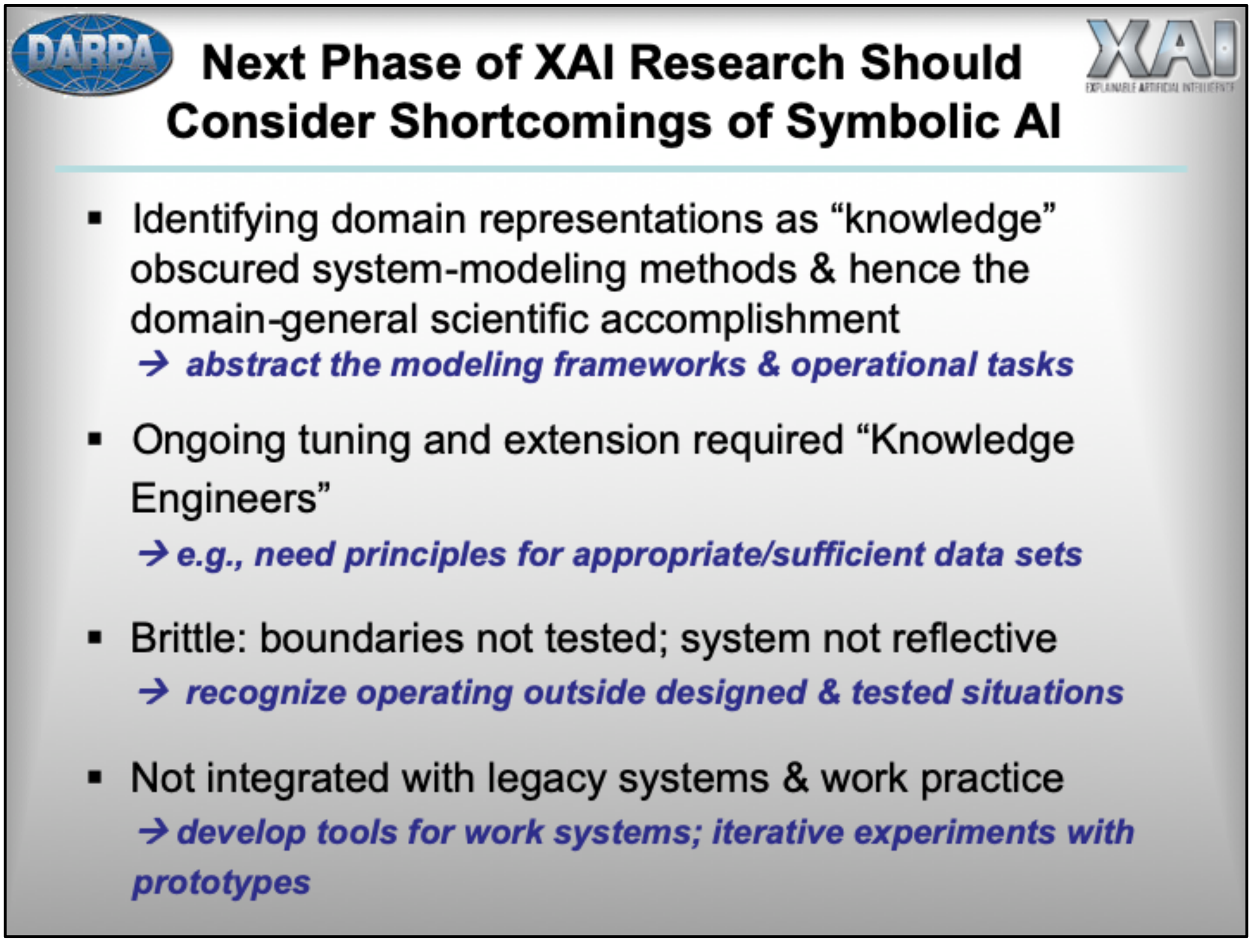
See http://Bill.Clancey.name for all publications 1977–present.

You can find all of my publications at my web site – it includes the explanation research of the 1970s, syntheses of the modeling methods of AI programming, and using Brahms simulation for work system design and agent systems.

The collection by Greenbaum and Kyng is a useful introduction to work system design. I also recommend Mindell’s analysis of how people interact with advanced automated systems in extreme environments.

For the Brahms work I particularly want to acknowledge Maarten Sierhuis. We worked closely with the anthropologists, Pat Sachs and Gitti Jordan.

In *Working on Mars* I present the MER robotic laboratories as collaboration tools for doing field science on Mars. In *Creative Engineering* I present the work of John Arnold, one of the pioneers of design thinking, placing it in the context of 1950s human factors perspectives



Next Phase of XAI Research Should Consider Shortcomings of Symbolic AI

- Identifying domain representations as "knowledge" obscured system-modeling methods & hence the domain-general scientific accomplishment
  → abstract the modeling frameworks & operational tasks
- Ongoing tuning and extension required "Knowledge Engineers"
  → e.g., need principles for appropriate/sufficient data sets
- Brittle: boundaries not tested; system not reflective
  → recognize operating outside designed & tested situations
- Not integrated with legacy systems & work practice
  → develop tools for work systems; iterative experiments with prototypes

ADDITIONAL SLIDE FOR Q&A

I'd also like to highlight some lessons learned from AI's Symbolic era that might be important for the success of today's "machine learning" research.

My intention is not to criticize DARPA's XAI projects or research program, but rather to put it in the context of descriptions and methods I have found useful in developing practical tools.

I focused on the last point in this presentation. Regarding the first point, see *Situated Cognition: On Human Knowledge and Computer Representations (Cambridge, 1997)* and the references listed on the prior page.