AI Systems and Trust:
Past, Present, and Future

Peter Friedland
Systems Plus/AFOSR

October 14, 2020

Issues to Consider

- Do we mean the same thing by AI Systems and Autonomous Systems?
  - AI: an artificial agent exhibiting some aspect of what we regard as human-like intelligence
  - Autonomous: an artificial agent capable of acting independently in a dynamic environment
- Truly autonomous operations vs. semi-autonomous vs. as part of multi-agent teams
- Adaptation vs. Learning
- Tools vs. Partners
- Trustworthiness vs. Trust
- Verification vs. Validation vs. other trust factors

The First Generation of Fielded AI Systems

1970-1995 (but actually continuing through today)

- **Symbolic methods—Knowledge-Based (“Expert”) Systems**
- **Heuristic vs. Algorithmic software so could not be traditionally verified**
- **Trust came from provenance—who were the experts?**
- **Reasoning usually transparent—explanation normally straightforward**
- **But the systems were “brittle” leading to the Second AI Winter**

7 JANUARY 1987

AIR FORCE
SYSTEMS COMMAND
ARTIFICIAL INTELLIGENCE
RESEARCH AND DEVELOPMENT
INVESTMENT PLAN

DCS SCIENCE & TECHNOLOGY

HEADQUARTERS AIR FORCE SYSTEMS COMMAND

Note:

FOR OFFICIAL USE ONLY – Cannot be distributed to contractors

SYSTEMS AUTONOMY TECHNOLOGY PROGRAM PLAN
EXECUTIVE SUMMARY

Ames Research Center
National Aeronautics and Space Administration
Moffett Field, California 94035

November 1987

AMES RESEARCH CENTER
GODDARD SPACE FLIGHT CENTER
JET PROPULSION LABORATORY
JOHNSON SPACE CENTER

KENNEDY SPACE CENTER
LANGLEY RESEARCH CENTER
LEWIS RESEARCH CENTER
MARSHALL SPACE FLIGHT CENTER

The Second (Current) Generation of Fielded AI Systems

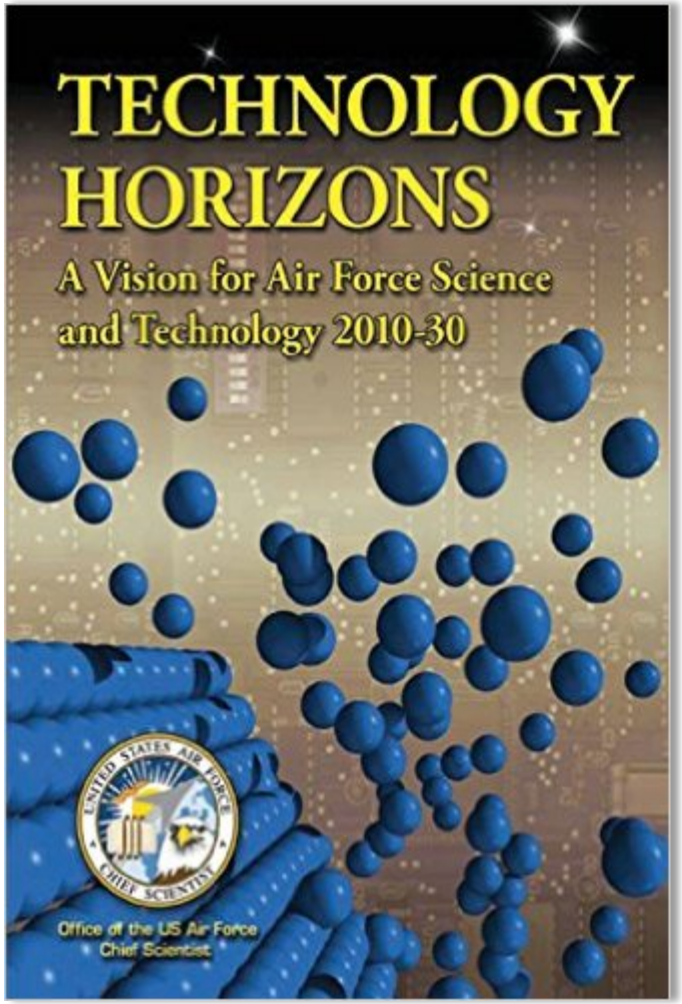
2005-Present

- **Statistical methods that apply an old technology: Neural Net ("Deep") Learning**
  - **Enabled by vast amounts of data and specialized hardware (GPUs and TPUs)**
  - **But the systems are opaque and trust through explanation difficult to impossible**

- **An additional focus on human-machine joint problem-solving**
  - **Trust here a complex multi-disciplinary, multi-cultural issue**

Technology Horizons

(Werner Dahm, Chief Scientist, USAF – 2010)


TECHNOLOGY HORIZONS
A Vision for Air Force Science and Technology 2010-30
UNITED STATES AIR FORCE CHIEF SCIENTIST
Office of the US Air Force Chief Scientist

"*Methods to enable trust* in highly adaptive autonomous systems *are a game-changing technology* needed to reap the enormous capability and cost benefits that such systems can offer, and to avoid the asymmetric advantage that adversaries could otherwise have with such systems."

"As autonomous technologies are fielded, their operational acceptance and success will largely depend upon calibrated operator trust..."

AFRL Autonomy S&T Strategy

(Maj Gen Massiello, AFRL Commander - 2013)

AIR FORCE RESEARCH LABORATORY
AUTONOMY
SCIENCE AND TECHNOLOGY
STRATEGY

THOMAS J. MASIELLO
Major General, USAF
Commander

DISTRIBUTION A. Approved for public release: distribution unlimited. (88ABW-2013-5023, 03 Dec 2013)

1

<u>Autonomy S&T vision</u>: Intelligent machines seamlessly integrated with humans - maximizing mission performance in complex and contested environments

Goal #1: Deliver flexible autonomy systems with highly effective *human-machine teaming*

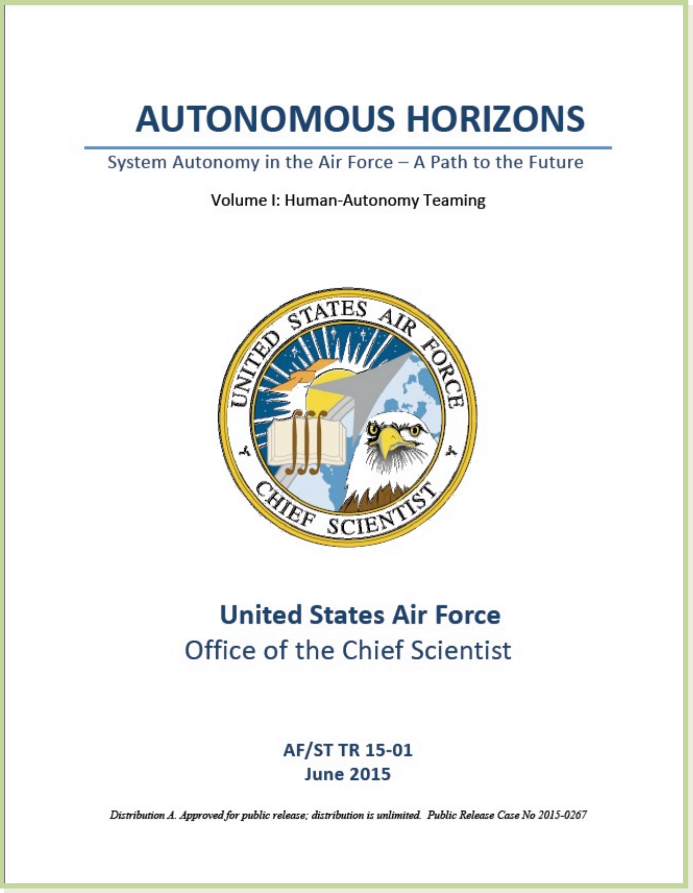
1.1. Enable & calibrate trust between human and machines:

https://www.youtube.com/watch?v=FXjrznpTTDs

Autonomous Horizons

Volume I: Human-Autonomy Teaming

(Mica Endlsey, Chief Scientist, USAF – 2015)



AUTONOMOUS HORIZONS

System Autonomy in the Air Force – A Path to the Future

Volume I: Human-Autonomy Teaming

United States Air Force
Office of the Chief Scientist

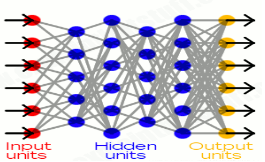
AF/ST TR 15-01
June 2015

Distribution A. Approved for public release; distribution is unlimited. Public Release Case No 2015-0267

“Particular care will need to be taken to allow airmen to develop trust that is well informed so that they will know how much to trust the autonomous system for a particular task, at a particular time, for the particular situation.”

Explainable AI – What Are We Trying To Do?

AI System



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

Watson

©IBM

AlphaGo

©Marcin Bajer/Flickr

Sensemaking

©NASA.gov

Operations

©Eric Keenan, U.S. Marine Corps

User

- 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?

Dramatic success in machine learning has led to an explosion of AI applications. Researchers have developed new AI capabilities for a wide variety of tasks. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems will be limited by the machine's inability to explain its thoughts and actions to human users. Explainable AI will be essential, if users are to understand, trust, and effectively manage this emerging generation of artificially intelligent partners.

The AFOSR Program in Basic Research into Human-Machine Trust

- Motivated by the Need for Humans to Trust Intelligent Systems Capable of Reacting to and Learning from their Environment and other Agents (both Human and Machine)—Reaction of the then AFOSR Chief Scientist, Tom Hussey, to the AFOSR AI Research Program
- First Workshop to Define the Program Held in Ocala, Florida January 31 to February 2, 2013
- Second Workshop to Extend the Program Held In Los Angeles, August 19-21, 2015
- Led to Conversion of an Existing Program in Trust and Influence among Humans into a Program focused mainly on Human-Machine Teams
- Over 25 Worldwide Activities at Universities and Government Research Laboratories have been Funded involving 10 disciplines and 15 countries
- Current AFOSR Program Officer is Dr. Laura Steckman

First Workshop on Human-Machine Trust

- Participation from Computer Scientists, Psychologists, Cognitive Scientists, Philosophers, and Human Factors Experts

- Kickoff Talk from AF Chief Scientist, Mark Maybury

- Themes:
  - Earning and Maintaining Trust
  - Robots, Cyber-Physical Systems, and Agents
  - Assessing Trust, Trustworthiness, and Appropriate Reliance
  - Adaptation and Emergent Behavior
  - Synthesis—Research Challenges

Results: Research Areas for AFOSR to Support

- **The Role of Human Predisposition**
  - **What are the Differential Impacts of Individual Human Characteristics (expertise, age, gender, etc.)?**
  - **What is the Role of Cognitive Workload/Stress in Shaping Trust?**
  - **Can the Machine be Trained to Recognize Human Trustworthiness Perception?**

- **Trust and Reciprocity**
  - **How do the Mechanisms of Human Interpersonal Trust Apply to Human-Machine Trust?**
  - **To What Degree do Humans Attribute the Traits of Trustworthiness to Machines?**
  - **Is an Attribution of Volition to an Autonomous System Necessary for Trust?**
  - **Is it Necessary for a Machine to "Have Something at Risk" for a Human to Trust it?**

- **Situational Factors (Tasks and State-of-Progress, Bystanders, Cultural Issues, etc.)**
  - **What Contextual Factors Shape Trustworthiness and Trust?**
  - **What is the Role of Public Attitudes in Shaping Individual Trust?**

Results: Research Areas for AFOSR to Support (cont.)

- Social Interaction
  - What Surface Cues of the Machine Trigger Trustworthiness (or Lack of Same)
  - What is the Role of Socially-Driven Communications Channels?
  - What Depth Cues over Time Foster and Maintain Perceptions of Trustworthiness?
  - What Level of Fidelity in Communication is Needed to Achieve Human Trust?
  - What is the Role of Confessions of Error and How Best to Convey that Error Information?

- Collaboration and Initiative
  - In Building Trust, When is it Important for the Machine to Take the Initiative?
  - What Form of Human-Machine Communications Enables "Smooth" Transfer of Control?
  - What Level of Visibility into the State of the Machine is Necessary?

Second Workshop on Human-Machine Trust

- **Participation from Computer Scientists, Psychologists, Cognitive Scientists, Philosophers, and Human Factors Experts**
- **Themes:**
  - **The Structure of Human-Machine Teams (are they like human-human teams or fundamentally different, how to distribute roles and skills)**
  - **Trust Dynamics: Calibration, Adaptation, and Repair**
  - **Distinguishing automation from autonomy (moving from tool to partner)**
  - **Anthropomorphism and Interpersonal Trust**
  - **Testbeds and Standards**
  - **Synthesis—Research Challenges**

Results: New Research Areas for AFOSR to Support

- **Human-Machine Team Resilience**
  - **What Characteristics of Human Team Resilience Carry over to Human-Machine Teams**
  - **Team Resilience as Opposed to Individual Resilience**
- **Machine Emotional Intelligence**
  - **Should a Machine show Empathy, Fear, and/or Uncertainty?**
  - **Should a Machine Show Self-Awareness of its own Errors and Apologize?**
- **The Costs of Anthropomorphism**
  - **When Positive and When Negative?**

The Third (Future) Generation of AI Systems

- **The Synthesis of the Symbolic and the Statistical Foundations of AI**

- **DARPA's AI Next Program: "DARPA envisions a future in which machines are more than just tools that execute human-programmed rules or generalize from human-curated data sets. Rather, the machines DARPA envisions will function more as colleagues than as tools. Towards this end, DARPA research and development in human-machine symbiosis sets a goal to partner with machines. Incorporating these technologies in military systems that collaborate with warfighters will facilitate better decisions in complex, time-critical, battlefield environments; enable a shared understanding of massive, incomplete, and contradictory information; and empower unmanned systems to perform critical missions safely and with high degrees of autonomy. DARPA is focusing its investments on a third wave of AI that brings forth machines that understand and reason in context."**

- **"Neural nets, and reinforcement learning and the things around that, are very big and being applied in many places, but they're not robust yet. Training these things and retraining them, for example, is still a bit of black art," he explained. "Formal verification — with the notion of flying in or trusting your life to decisions being made by a system autonomously — today, I wouldn't do it without having humans involved, without having controllers." Peter Highnam, DARPA Deputy Director**