U.S. AIR FORCE

USSF

AFRL

# Autonomy Capability for Next Generation Air and Space Dominance

Dr. Kevin “Nagel” Schmidt

AFRL Autonomy Capability Team (ACT3)

# Agenda

- What is AFRL's ACT3
- Applied State-of-the-art AI to USAF and USSF challenges
  - Autonomous Air Combat Operations (AACO)
  - Space Autonomy
  - Air/Space Safe, Ethical Autonomy
- Next Generation AI
  - Disruptive Capabilities = Cutting Edge Research & Development
  - QuEST = Qualia Exploitation of Sensing Technology

# What is AFRL's Autonomy Capability Team (ACT3)?

- Applying industry best practices to **scale AI solutions** across the DAF (USAF/USSF)
- Organic AI **expertise from across AFRL** enterprise, growing group since 2017
- Out of the box solutions from our people **outside the gate**
- Dev/Ops and Research in **custom IT environment**
- Integrated Civ, Ctr, Mil Teams with focus, **technical depth**

Autonomous Horizons v2: https://www.airuniversity.af.edu/AUPress/Display/Article/1787830/autonomous-horizons-the-way-forward

# Air/Space Force Cognitive Engine

A NEW business model for invention / development / fielding of AI

Execute AI pilot projects to gain momentum

Develop an AI / Analytics Strategy

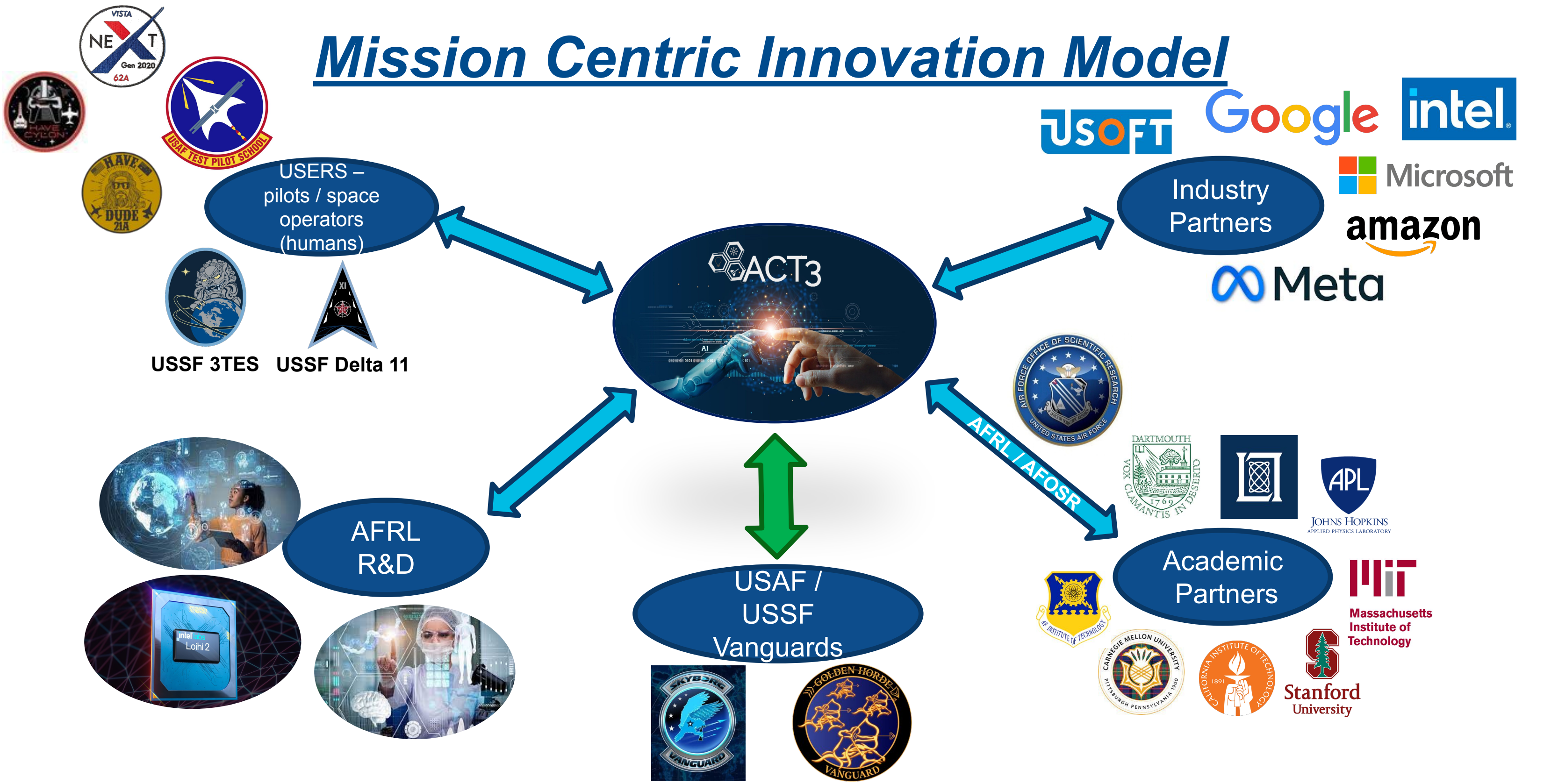
Infrastructure: build an in-house AI / Analytics Team

Develop internal and external communications

Provide broad AI / Analytics Training

**Applying industry best practices to scale AI solutions across the AF/USSF**

# Mission Centric Innovation Model




## User-Producer Innovation with the Art of the Possible

# Applying the State-of-the-Art AI to USAF/USSF Challenges

# Significant Achievements of Reinforcement Learning

RL has already shown superhuman performance in several big “human versus machine” competitions - high dimensional state spaces, partial observability, complex strategy

- AlphaGo - Deepmind trained agents to play GO (2016)
- OpenAI 5 - OpenAI trained agents to play DOTA 2 (2019)
- AlphaStar - Deepmind trained agents to play StarCraft 2 (2019)
- Alpha Dog Fight Trials - Heron Systems trained agents to aerial dogfight (2020)
- GPT-4 passes Bar Exam, US Medical Licensing Exam, college course exams (2023)

Photo Courtesy of DARPA

# Autonomous Air Combat Operations

***Vision Statement:*** AFRL ACT3 Autonomous Air Combat Operation (AACO) was formed to Demonstrate AI Tactical Autopilot engaging in multi-ship / multirole Beyond Visual Range (BVR) & ISR combat operations with tactical proficiency



**AACO Formed** — FY19Q2

**Learjet** — FY21/FY22
**25+ sorties**
**Aviate, Navigate, BVR, LVC**

**X62 VISTA (F16)** — FY23Q1
**8 sorties**
**Aviate, Navigate, BVR, LVC**

**XQ-58 Valkyrie** — FY23Q2
**Aviate, Navigate, BVR, LVC**
**Multi ship**

**Impacts**

Generating unprecedented data sets to support future AF needs

*AACO 100% government owned tools, agents, and capabilities are the base of DARPA AIR*

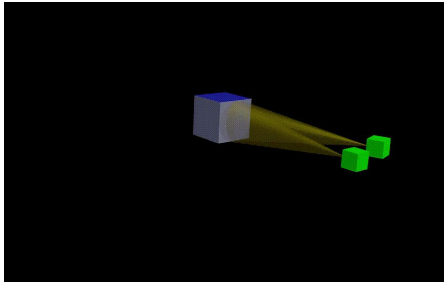
# Current Core Focus: Cooperating Space

Quickly react, plan, and decide on appropriate courses of action for inspection tasking in proximity operations in support of In-space Servicing, Assembly, and Manufacturing (ISAM).

For the following design reference missions (DRMs):

**DRM1**. Chief spacecraft provides sensing and state estimation for multiple deputy spacecraft.

**DRM2**. Multiple deputy spacecraft collaborate to inspect a chief spacecraft.



With the following high-level General Technical Objectives (GTOs):

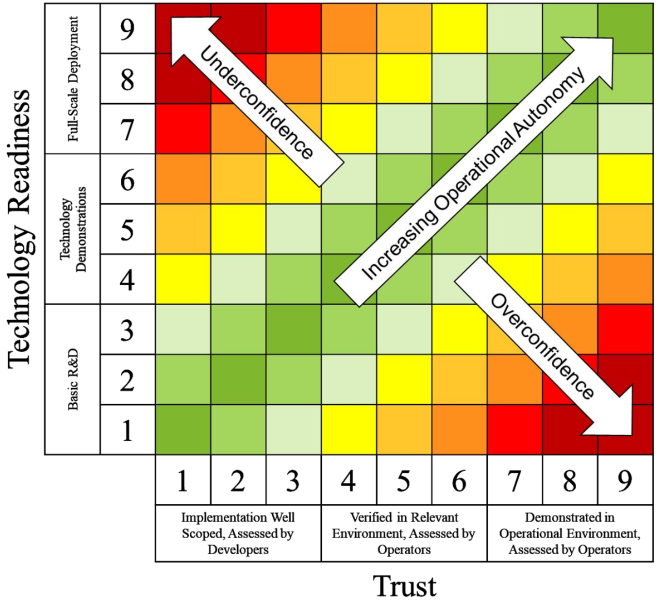
**GTO1**. Develop flexible **human-autonomy interfaces** that accept operator-based mission preferences and provide course of action visualization, comparison, and selection mechanisms

**GTO2**. Develop **reinforcement learning-based neural network multi-agent controllers** that incorporate mission- and task-specific operator preferences in identification of courses of action

**GTO3**. Develop **run time assurance** approaches that mitigate hazards and allow the autonomy to stay on mission for longer in the face of unexplored system parameters, scenarios, and/or poorly modeled aspects of the system.

# Space Trusted Autonomy Levels

- Active Partner in the NRO/NASA/USSF Space Trusted Autonomy group under the Space Science and Technology Partnership forum
- Defined ***Space Trusted Autonomy Readiness (STAR) Levels***
  - Defines **capability** and **trust** levels
  - Includes background on trust in automation research
  - **Consistent with other forms of readiness levels**: original TRLs, algorithm RLs, manufacturing RL, commercialization RL, data RL, machine learning RL, etc.




Kerianne L. Hobbs, Joseph B. Lyons, Martin S. Feather, Benjamin P Bycroft, Sean Phillips, Michelle Simon, Mark Harter, Kenneth Costello, Philip C. Slingerland, Yuri Gawdiak, Stephen Paine, “Space Trusted Autonomy Readiness Levels,” IEEE Aerospace, Big Sky, MT, March 4-11, 2023. Preprint: https://arxiv.org/pdf/2210.09059

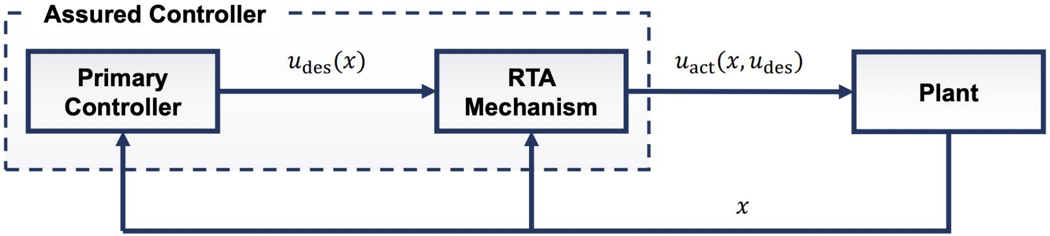
# Ethics, Safety and Trust in AI

- There is no such thing as Ethical AI, only Ethical use of AI
    - ***Trust*** is defined as a willingness to accept vulnerability in situations characterized by uncertainty.
    - ***Safety*** is freedom from harm during operations.
    - ***Ethics*** are rules created by societies and cultures governing moral and just usage of a technology.
- Ethical considerations begin at design time and continue through operational use
- Potential to use Model Cards or Datasheets for Datasets as communication of ethics from design to use

[Draft] Joseph B. Lyons, Kerianne Hobbs, Steve "Cap" Rogers, Scott H. Clouse, "Responsible (Use of) AI," in Understanding the Role of Humanity in Responsible Deployment of Intelligent Technologies in Socio-technical Ecosystems, Special Issue of Frontiers in Neuroergonomics, 2023.
[In Draft] Kerianne Hobbs, Bernard Li, "Safety, Trust, and Ethics Considerations for Human-AI Teaming in Aerospace Control," AIAA SciTech, 8–12 January 2024, Orlando, FL.

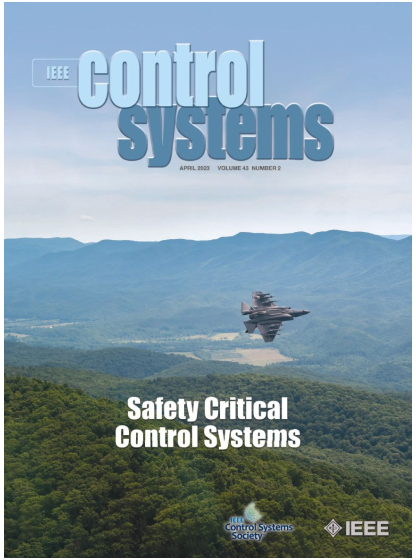
# Verifying and Guaranteeing Safety

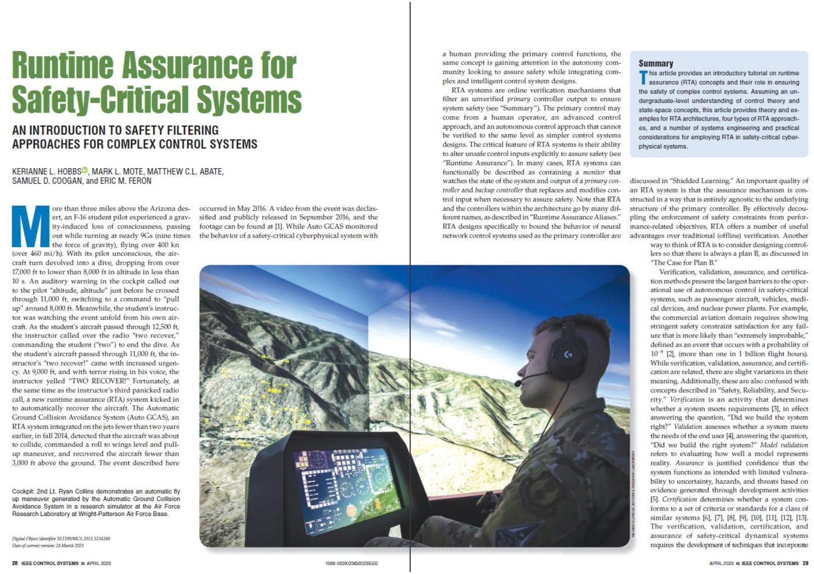
- Run Time Assurance
- Safe RL, Rigorous Evaluation
- Neural Network Verification
- Hazard Analysis
- Assurance Cases




Selected Publications

1. Kerianne L. Hobbs, Mark L. Mote, Matthew C.L. Abate, Samuel D. Coogan, and Eric M. Feron "**Run Time Assurance for Safety Critical Systems:** An Introduction to Safety Filtering Approaches for Complex Control Systems," IEEE Control Systems Magazine. April 2023. Preprint: https://arxiv.org/pdf/2110.03506.pdf
2. [Submitted] Jonathan Rowanhill, Ashlie B. Hocking, Aditya Zutshi, Kerianne L. Hobbs "Conformance of Run-Time Assurance to MIL-HDBK-516C through Verified System **Arguments**" Digital Avionics Systems Conference 2023.
3. Hobbs, K., Heiner, B., Busse, L., Rowanhill, J., Hocking, A. B., Zutshi, A., "Systems Theoretic Process Analysis of a Run Time Assured Neural Network Control System," AIAA SciTech, National Harbor, MD, Jan. 2023. Preprint: https://arxiv.org/pdf/2209.00552
4. Diego Manzanas Lopez, Hoang-Dung Tran, Taylor T. Johnson, Stanley Bak, Xin Chen, Kerianne L. Hobbs, "Evaluation of **Neural Network Verification Methods** for Air to Air Collision Avoidance" AIAA Journal of Air Transportation Systems, 2022.



Runtime Assurance for Safety-Critical Systems

AN INTRODUCTION TO SAFETY FILTERING APPROACHES FOR COMPLEX CONTROL SYSTEMS

KERIANNE L. HOBBS, MARK L. MOTE, MATTHEW C.L. ABATE, SAMUEL D. COOGAN, and ERIC M. FERON

More than three miles above the Arizona desert, an F-16 student pilot experienced a gravity-induced loss of consciousness, passing out while turning at nearly 9Gs (nine times the force of gravity), flying over 400 kn (over 460 mi/h). With its pilot unconscious, the aircraft turn devolved into a dive, dropping from over 17,000 ft to lower than 8,000 ft in altitude in less than 10 s. An auditory warning in the cockpit called out to the pilot "altitude, altitude" just before he crossed through 11,000 ft, switching to a command to "pull up" around 8,000 ft. Meanwhile, the student's instructor was watching the event unfold from his own aircraft. As the student's aircraft passed through 12,500 ft, the instructor called over the radio "two recover," commanding the student ("two") to end the dive. As the student's aircraft passed through 11,000 ft, the instructor's "two recover!" came with increased urgency. At 9,000 ft, and with terror rising in his voice, the instructor yelled "TWO RECOVER!" Fortunately, at the same time as the instructor's third panicked radio call, a new runtime assurance (RTA) system kicked in to automatically recover the aircraft. The Automatic Ground Collision Avoidance System (Auto GCAS), an RTA system integrated on the jets fewer than two years earlier, in fall 2014, detected that the aircraft was about to collide, commanded a roll to wings level and pull-up maneuver, and recovered the aircraft fewer than 3,000 ft above the ground. The event described here occurred in May 2016. A video from the event was declassified and publicly released in September 2016, and the footage can be found at [1]. While Auto GCAS monitored the behavior of a safety-critical cyberphysical system with a human providing the primary control functions, the same concept is gaining attention in the autonomy community looking to assure safety while integrating complex and intelligent control system designs.

RTA systems are online verification mechanisms that filter an unverified *primary* controller output to ensure system safety (see "Summary"). The primary control may come from a human operator, an advanced control approach, and an autonomous control approach that cannot be verified to the same level as simpler control systems designs. The critical feature of RTA systems is their ability to alter unsafe control inputs explicitly to assure safety (see "Runtime Assurance"). In many cases, RTA systems can functionally be described as containing a *monitor* that watches the state of the system and output of a *primary controller* and *backup controller* that replaces and modifies control input when necessary to assure safety. Note that RTA and the controllers within the architecture go by many different names, as described in "Runtime Assurance Aliases." RTA designs specifically to bound the behavior of neural network control systems used as the primary controller are discussed in "Shielded Learning." An important quality of an RTA system is that the assurance mechanism is constructed in a way that is entirely agnostic to the underlying structure of the primary controller. By effectively decoupling the enforcement of safety constraints from performance-related objectives, RTA offers a number of useful advantages over traditional (offline) verification. Another way to think of RTA is to consider designing controllers so that there is always a plan B, as discussed in "The Case for Plan B."

Verification, validation, assurance, and certification methods present the largest barriers to the operational use of autonomous control in safety-critical systems, such as passenger aircraft, vehicles, medical devices, and nuclear power plants. For example, the commercial aviation domain requires showing stringent safety constraint satisfaction for any failure that is more likely than "extremely improbable," defined as an event that occurs with a probability of $10^{-9}$ [2], (more than one in 1 billion flight hours). While verification, validation, assurance, and certification are related, there are slight variations in their meaning. Additionally, these are also confused with concepts described in "Safety, Reliability, and Security." *Verification* is an activity that determines whether a system meets requirements [3], in effect answering the question, "Did we build the system right?" *Validation* assesses whether a system meets the needs of the end user [4], answering the question, "Did we build the right system?" *Model validation* refers to evaluating how well a model represents reality. *Assurance* is justified confidence that the system functions as intended with limited vulnerability to uncertainty, hazards, and threats based on evidence generated through development activities [5]. *Certification* determines whether a system conforms to a set of criteria or standards for a class of similar systems [6], [7], [8], [9], [10], [11], [12], [13]. The verification, validation, certification, and assurance of safety-critical dynamical systems requires the development of techniques that incorporate


Summary

This article provides an introductory tutorial on runtime assurance (RTA) concepts and their role in ensuring the safety of complex control systems. Assuming an undergraduate-level understanding of control theory and state-space concepts, this article provides theory and examples for RTA architectures, four types of RTA approaches, and a number of systems engineering and practical considerations for employing RTA in safety-critical cyber-physical systems.


Cockpit: 2nd Lt. Ryan Collins demonstrates an automatic fly up maneuver generated by the Automatic Ground Collision Avoidance System in a research simulator at the Air Force Research Laboratory at Wright-Patterson Air Force Base.

RICHARD ELDRIDGE, AIR FORCE RESEARCH LABORATORY


Digital Object Identifier 10.1109/MCS.2023.3234380
Date of current version: 24 March 2023





1066-033X/23©2023IEEE

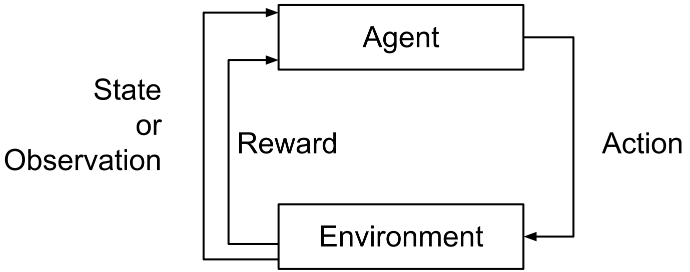
# Creating the Next Generation of AI

# Disruptive Capability Upgrades – Collaborative Combat Aircraft

**Speed and Range of Adaptation Determined by *Ability to Transfer Across Representations***

Conflict — Mission — Engagement

## Advancing Current Capabilities




- **Primary Problem #1: Efficient Agent Training**
- Catastrophic Interference
  - Effects of rewards and training curricula
  - Retraining components

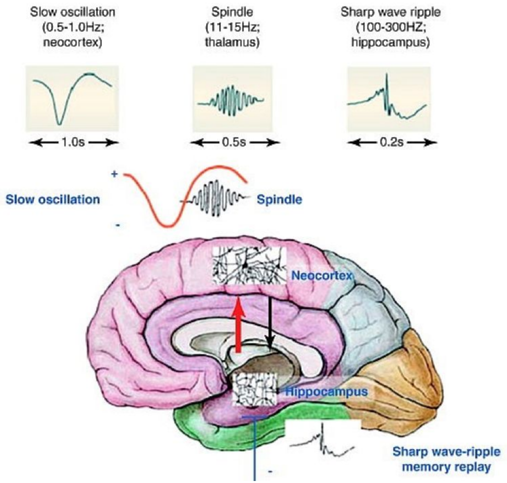
## Disrupting the Fight

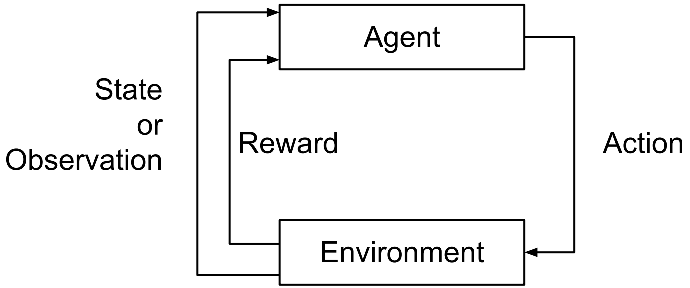

**Need:**

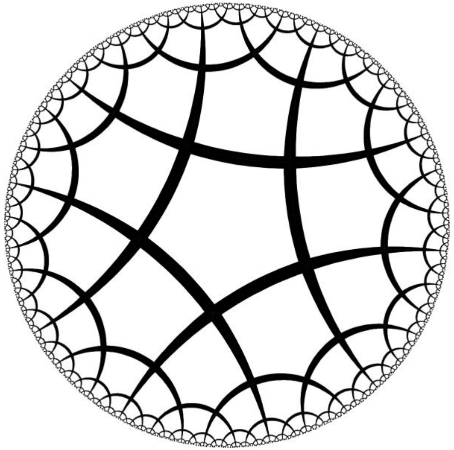
- Real-time decentralized & distributed learning
- Within-mission model updates
- Robustness to disrupted, disconnected, intermittent, and low-bandwidth (DDIL) environments

**User-Producer Innovation with the Art of the Possible**

# Why Consciousness?



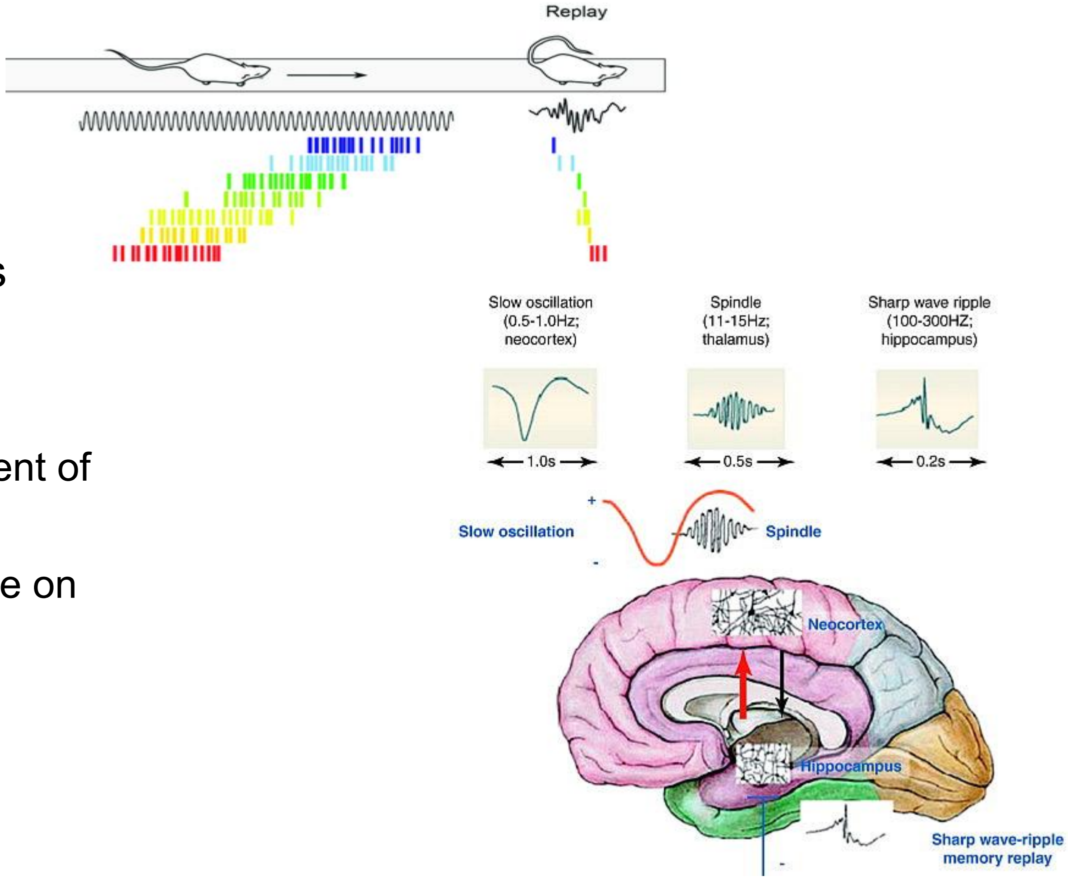



**Understand Human Cognition** **Build More Flexible Machines** **Study Interesting Structures**

**Representations of Computational Processes**

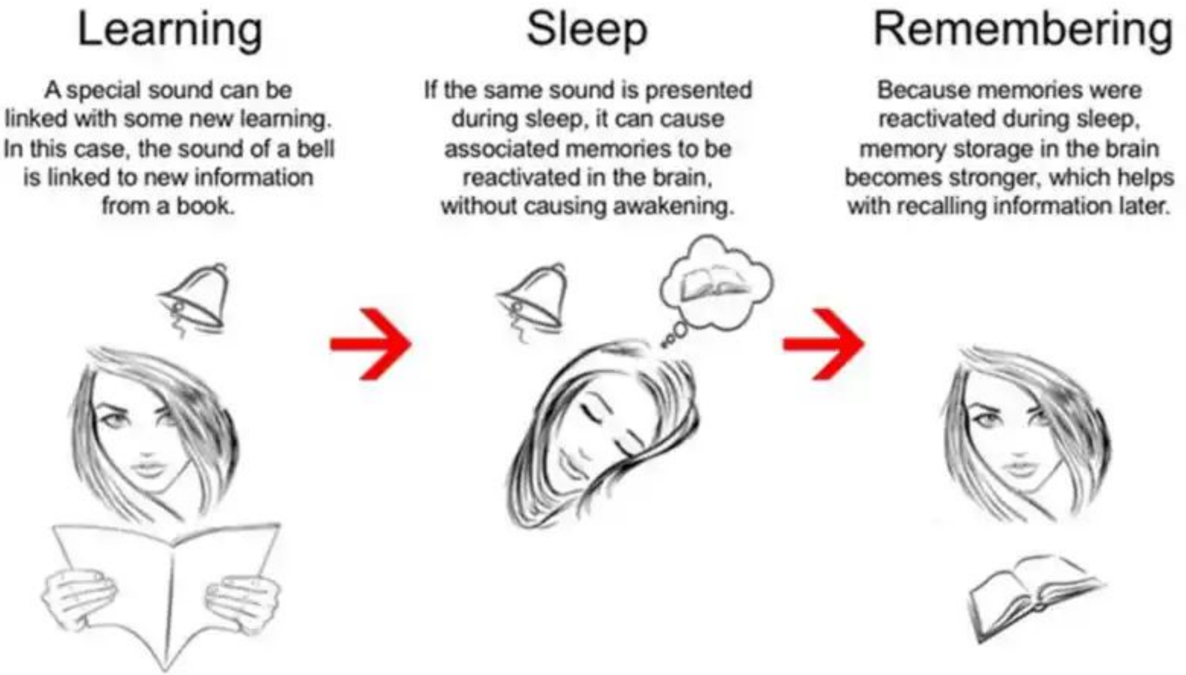
# Memory Consolidation




- Complementary Learning Systems
  - Memories first depend on the hippocampus
  - Hippocampus supports reinstatement of recent memories in the neocortex
  - Neocortical synapses change a little on each reinstatement
- ***Interleaved learning and catastrophic interference***

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

# Targeted Memory Reactivation




Schmidt, K., Larue, O., Kulhanek, R., Flaute, D., Veliche, R., Manasseh, C., Dellis, N., Culbertson, J., Clouse, H. & Rogers, S. (2023). Representational Tenets for Memory Athletics. arXiv preprint arXiv:2303.11944.

# Brain-Inspired AI


## System 1

- **Reinforcement-based mechanisms**
- Value of stimuli and actions are learned incrementally and through repeated experience
- Extracts statistical co-occurrences among features
- Neocortex
  - Slow acquisition of structure
  - Parametric
  - Efficient representations for generalization

## System 2

- **Conscious memory**
- Instance based mechanisms
- Allow experiences to be encoded rapidly (in "one shot")
- Hippocampus
  - Rapid storage: individual experiences
  - Non-parametric instance-based system
  - Sparse non-overlapping representations (poor generalization)

Schmidt, K., Culbertson, J., Cox, C., Clouse, H., Larue, O., Molineaux, M., Rogers, S. (2021). What is it Like to Be a Bot: Simulated, Situated, Structurally Coherent Qualia (S3Q) Theory of Consciousness. https://arxiv.org/abs/2103.12638

Han, C., Schmidt, K., Grandoit, E., Shu, P., McRobert, C., Reber, P. (2022). 'Cognitive Neuroscience of Implicit Learning: Implications for Complex Learning and Expertise, in Arthur S. Reber, and Rhianon Allen (eds), *The Cognitive Unconscious: The First Half Century*, New York, online edn, Oxford Academic.

**We need a suitable representation language**

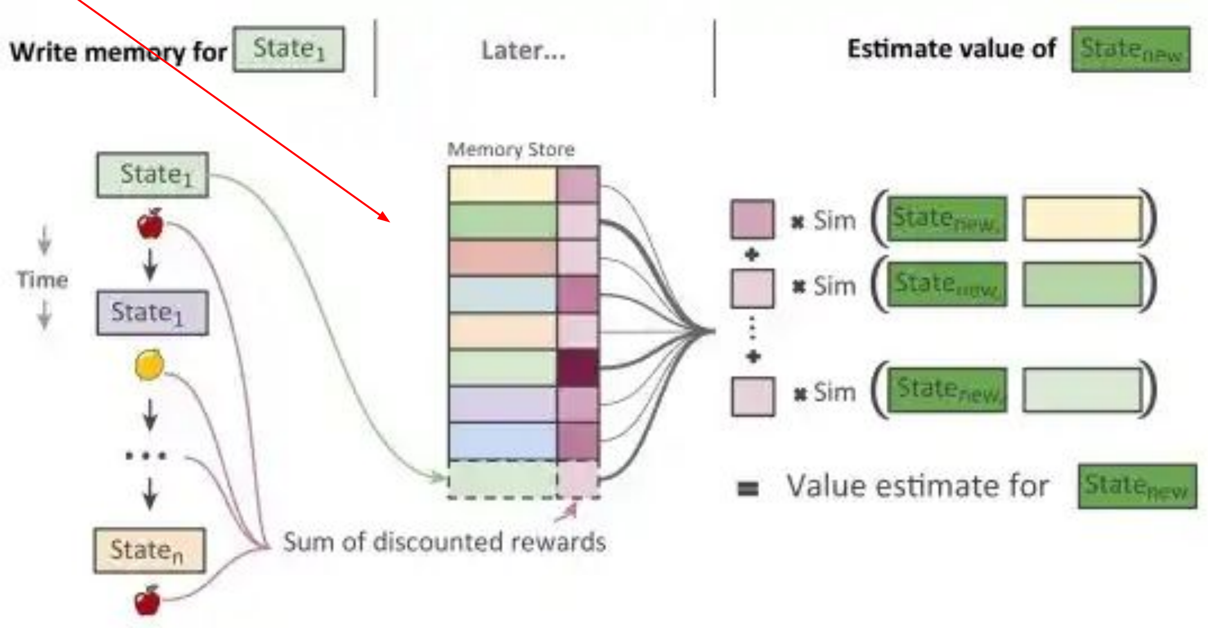


CellPress REVIEWS

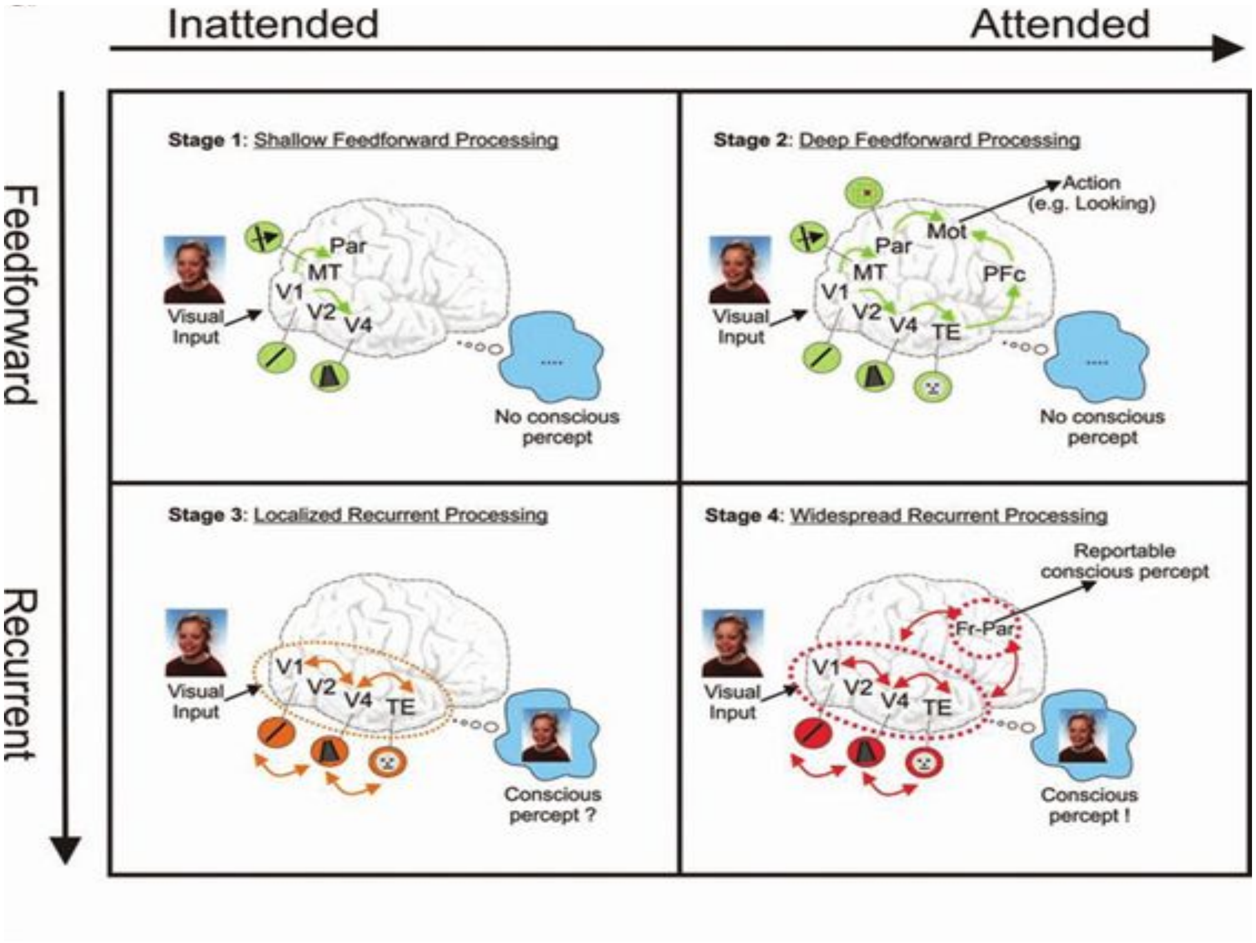
Review

# Reinforcement Learning, Fast and Slow

Matthew Botvinick,[1,2,*] Sam Ritter,[1,3] Jane X. Wang,[1] Zeb Kurth-Nelson,[1,2] Charles Blundell,[1] and Demis Hassabis[1,2]

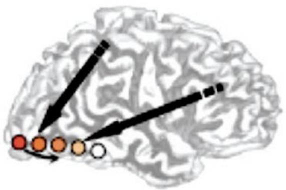


Feed forward universal function approximator cannot compute consciousness



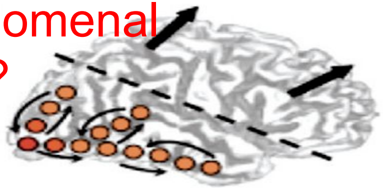

Attention is NOT all you need

Artificial Neural Network



- Subliminal Attended or Unattended
- Little Local Activity
- Feedforward
- No reportability

Hierarchical LSTM graphical NN? Better suited for phenomenal content?



- **Conscious??**
- Recurrent Activation confined to sensorimotor processing
- No reportability

Transformer, LLM? Well suited for language?

- Conscious
- Global Ignition / Gamma Synchrony / P3
- Durable Activation
- Reportable

# Questions?

Join us @

# QuEST

Fridays at 12p eastern!!

Qualia Exploitation of Sensing Technology

Public Seminar for the World's Best on the Topic


## Kevin Schmidt, PhD

ACT3 Senior Neuroscientist

Autonomy Capability Team (ACT3) | AFRL/RYZA
Sensors Directorate | Air Force Research Laboratory
Wright-Patterson AFB, OH

kevin.schmidt.15@us.af.mil


meet.google.com/dui-wwjj-fzr