Calibrated Trust vs Blind Faith

Testing AI/ML Powered Emerging Technologies for Aviation

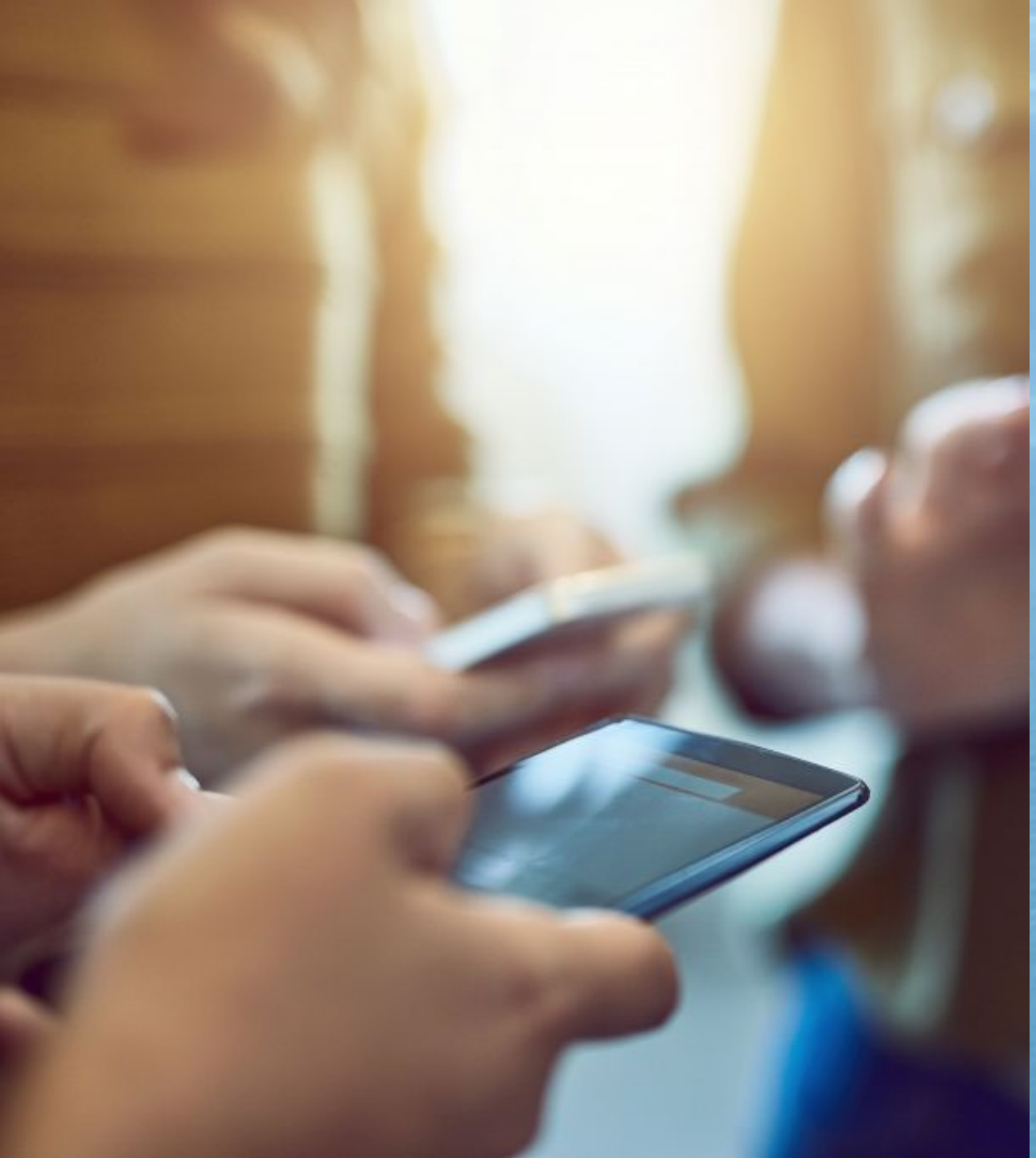
Col Dan "Animal" Javorsek, PhD

Commander, AFOTEC Det 6

Director, F-35 US Operational Test Team

21 September 2022


IBM
System 370

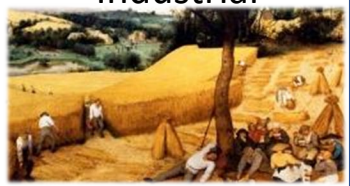


The Augmented Age is the next great era

Enabled by software defined systems augmenting human decisions

Information

Industrial



Agricultural

Hunter-gatherer

???
Decades
Centuries
Millennia
Millions of Years

Recent Contextual Adaptation Increases Combat Utility

Force Acquisition
thinks ‘procurement’

Mission Planner
thinks ‘planning’

Battle Manager
thinks ‘mission’

Platforms

Systems

History of AI

1960 1970 1980 1990 2000 2010 2020.....

AI Funding in Waves

1st AI WAVE

2nd AI WAVE

- Abstract logic
- Explicit representations

Handcrafted Knowledge (rules)

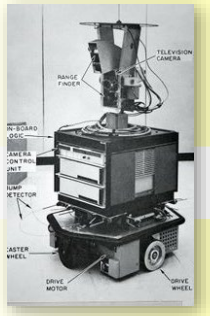
Statistical Learning (neural nets)

- Biological inspiration
- Implicit representations

3rd AI WAVE

Contextual Adaptation

- Explanatory models to explain decisions
- A contextual model can:
  - Reason
  - Perceive
  - Learn
  - Abstract



Perception & Robotics

https://i.kinja-img.com/gawker-media/image/upload/s--yohGERi5--/c_scale,f_auto,fl_progressive,q_80,w_800/199sapq7usql7jpg.jpg

Autonomous Land Vehicle

https://d253pvgap36xx8.cloudfront.net/editor_uploads/1277/2015/01/13/stanley1.jpg

Grand Challenge Autonomous Car

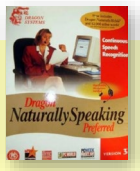
Problem Solving

Project on Mathematics and Computing (MAC)

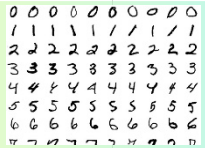
DART Logistics Planner

Formally Verified Flight Software

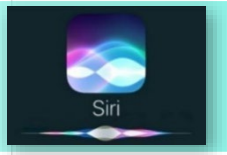
Natural Language Understanding



Speech Recognition



Handwriting

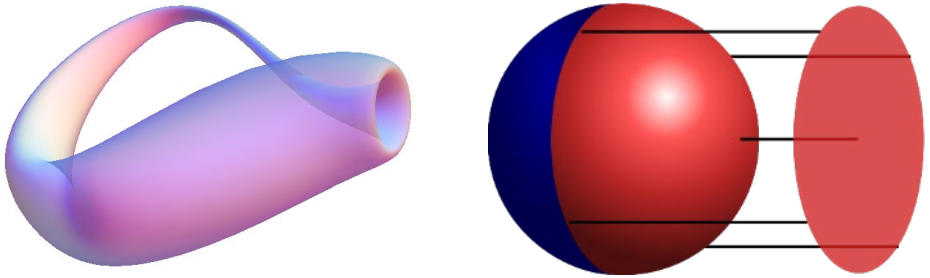


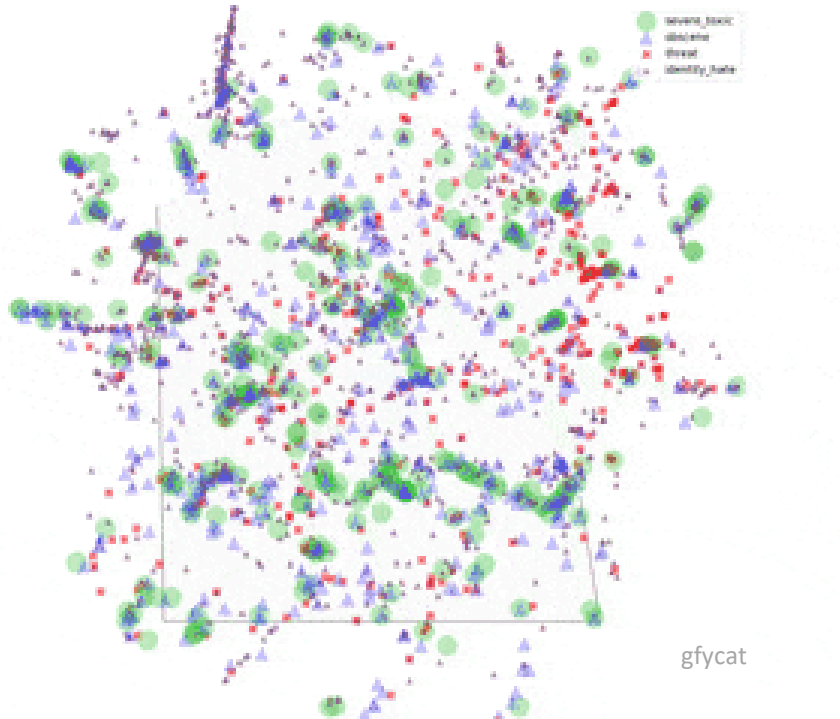
SIRI

1960 1970 1980 1990 2000 2010 2020.....

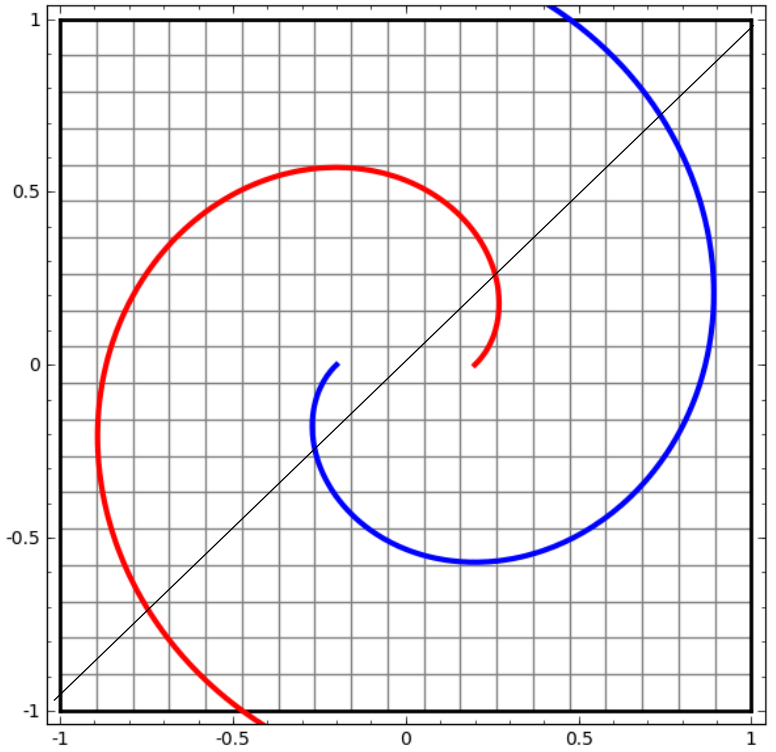
Assumptions when applying AI to air combat

- Air combat problems consist of two time constants
  - immediate tactic selection
  - long-term strategic impact
- The air combat problem is manifold bounded





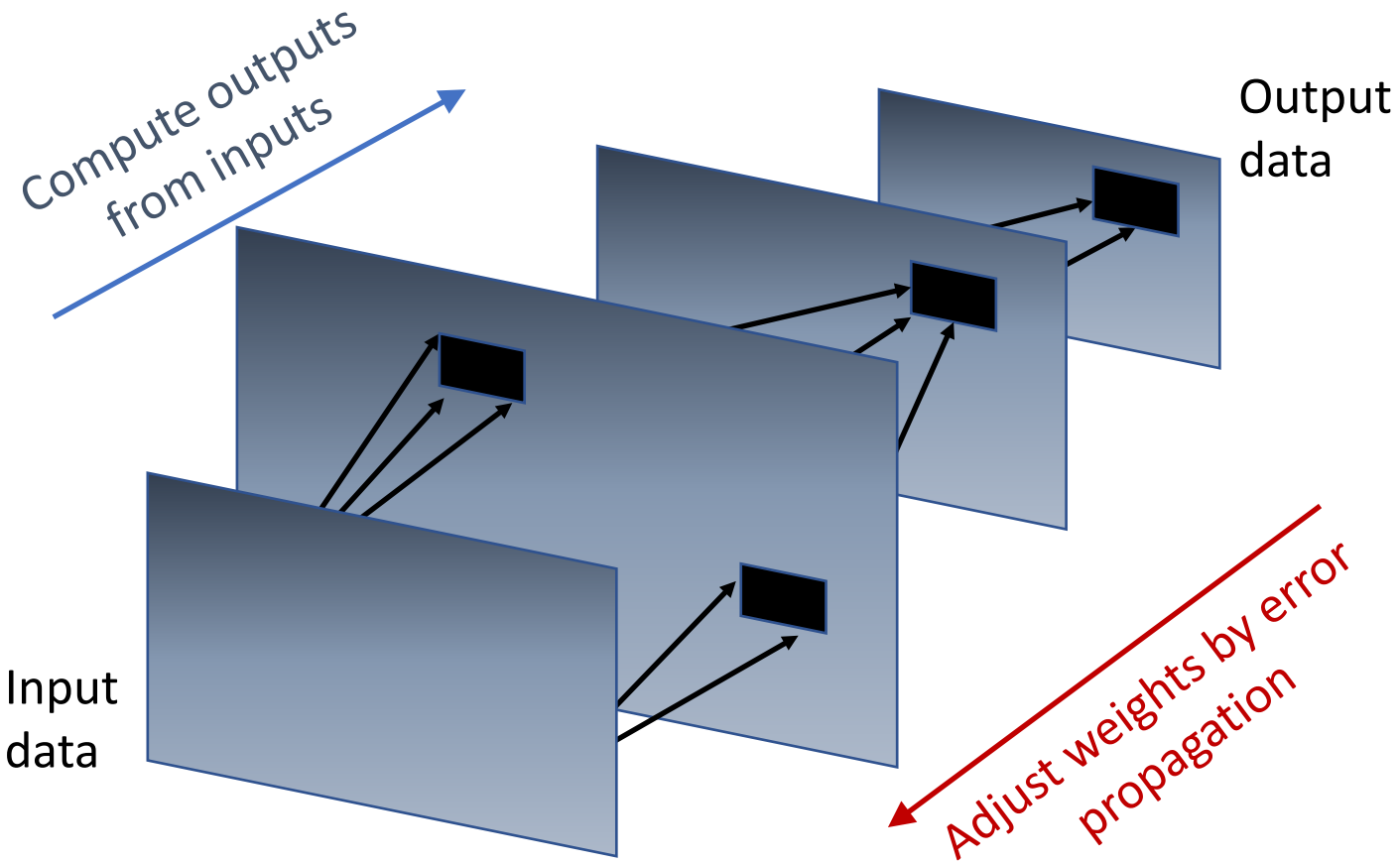
gfycat

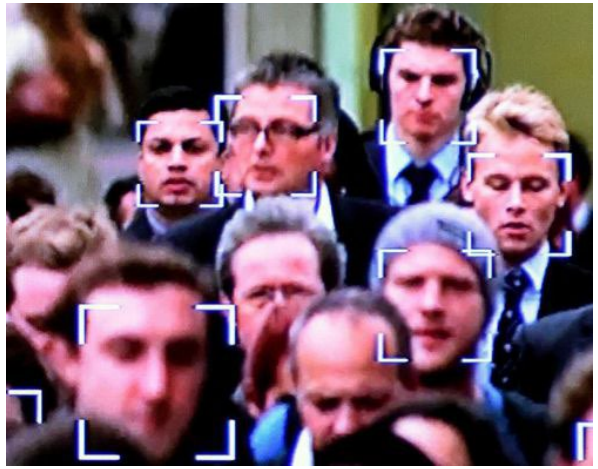


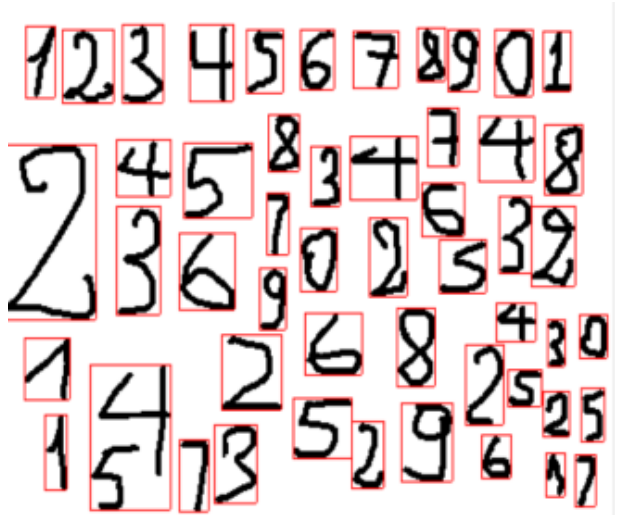
github

Neural nets are just spreadsheets on steroids

- Each layer stretches and squashes the data space until the manifolds are cleanly separated

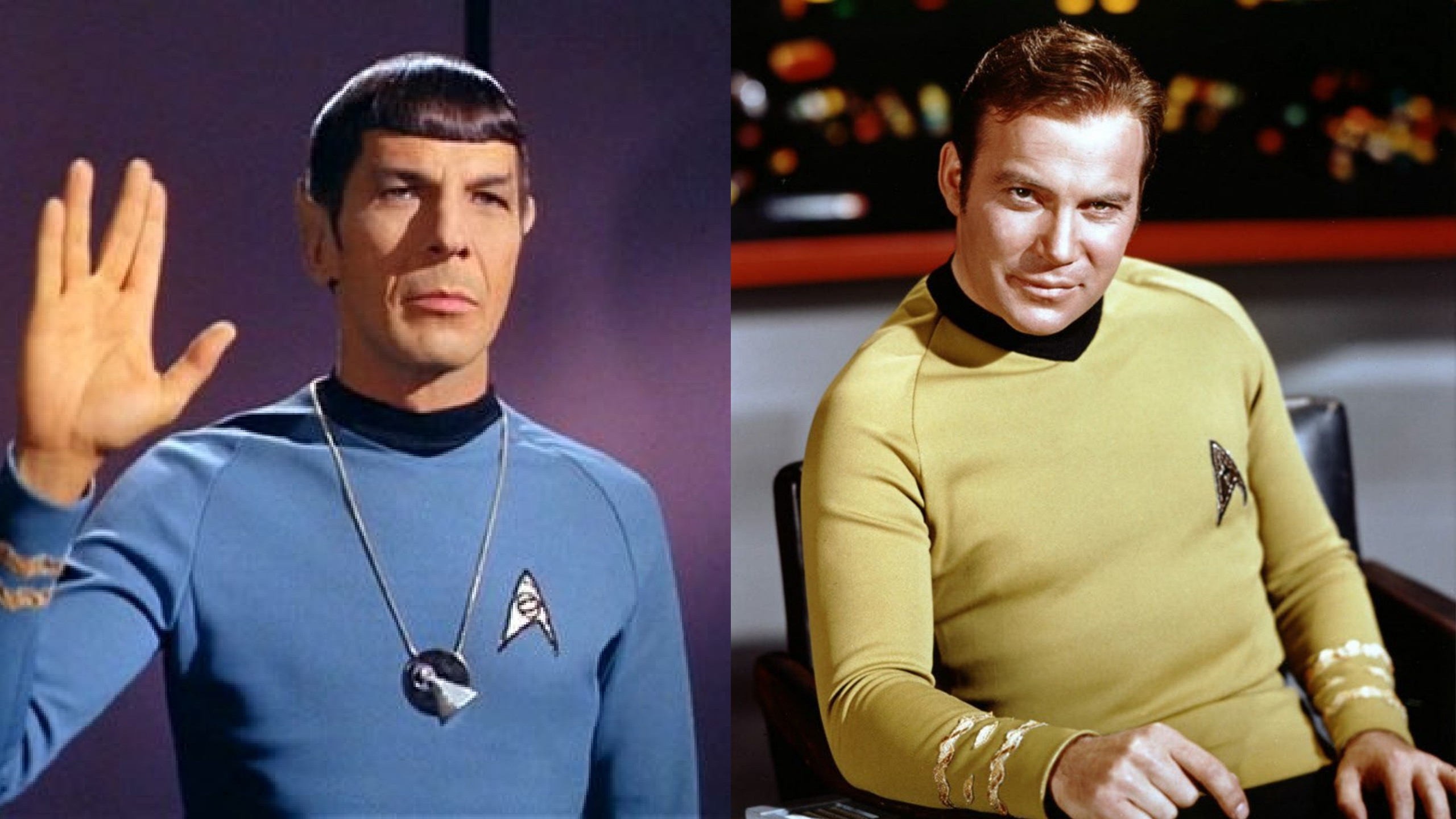


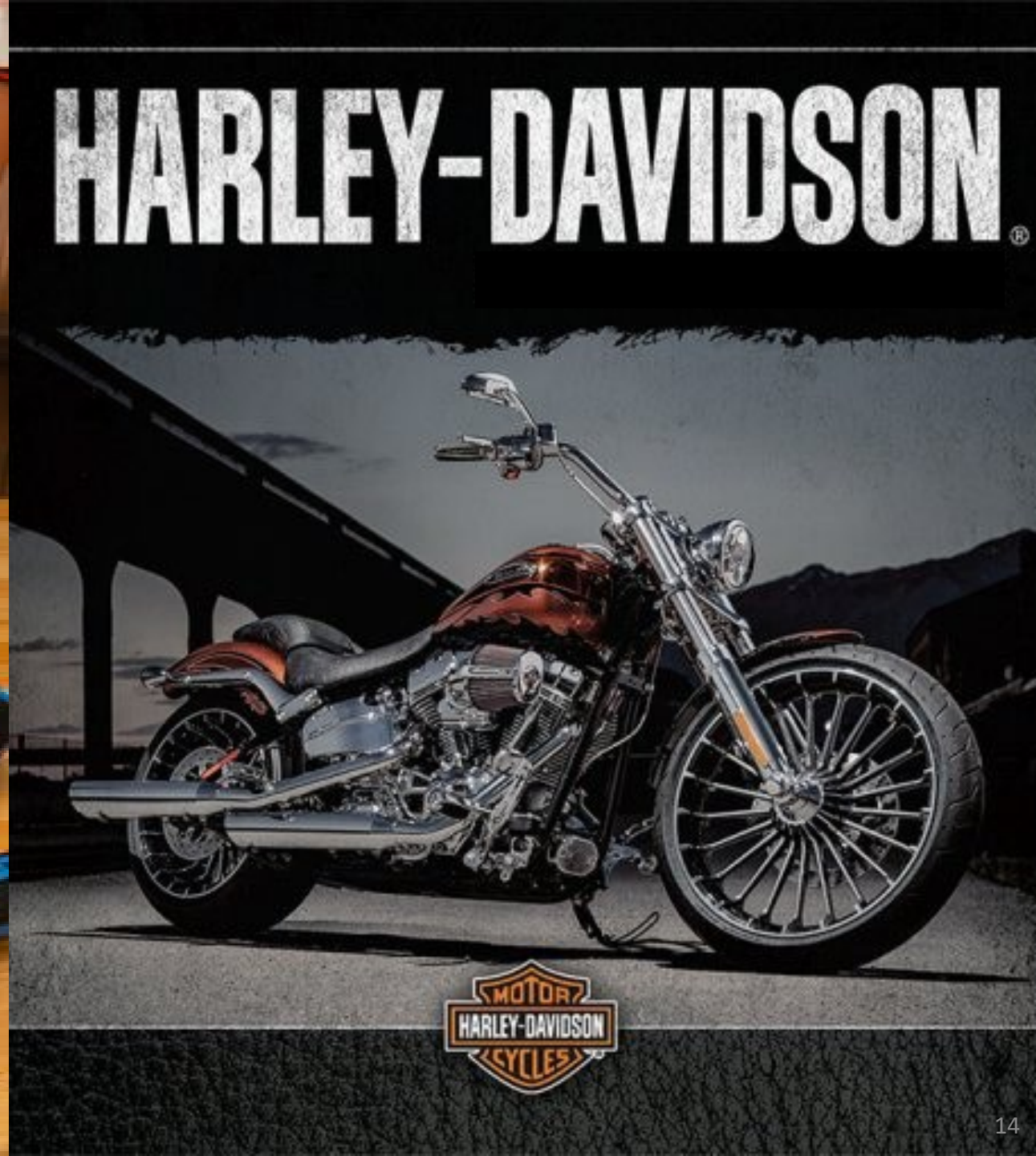




Evolution of the AI playground for strategy and decision making

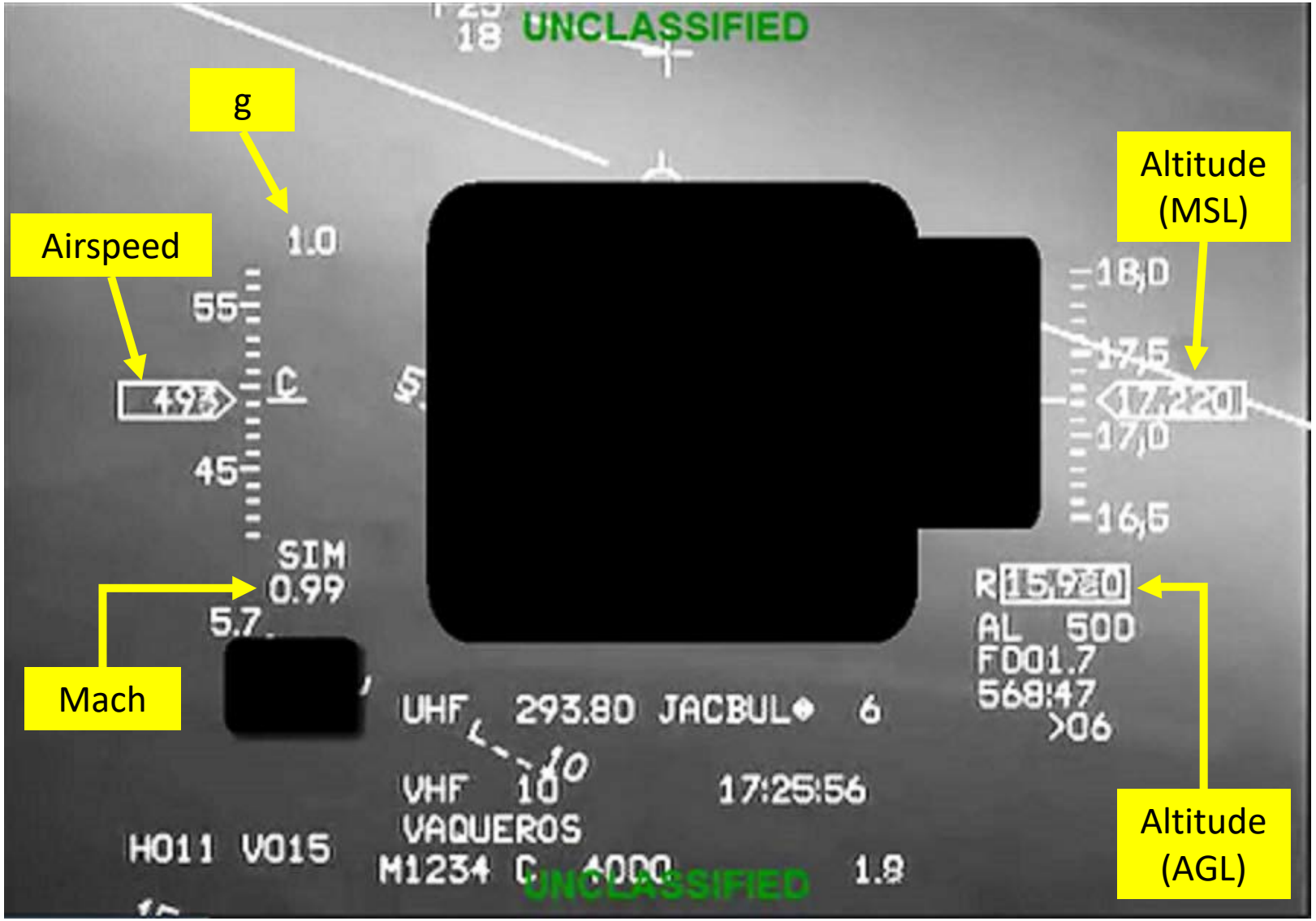
game/model				Mosaic ACE TA3, ACK and others
AI	IBM DEEP BLUE	AlphaGo	AlphaStar	AlphaDogfight AlphaMosaic
comment	Deep Blue beats 12-time world champion and grandmaster Garry Kasparov in 3 games to zero (with 3 draws)	AlphaGo beats 18-time world champion Lee Sedol in 4 games to one	AlphaStar beats TLO and MaNa (two of the best pro gamers in the world) 10-0	Once useful models are developed, AI will be applied to determine the best ways to beat the game
	 © AFP	 Lee Jin-man / Associated Press	 © Google Deepmind	 Ernest et al.
year	1997	2016	2019	2020+




HARLEY-DAVIDSON®
MOTOR
HARLEY-DAVIDSON
CYCLES

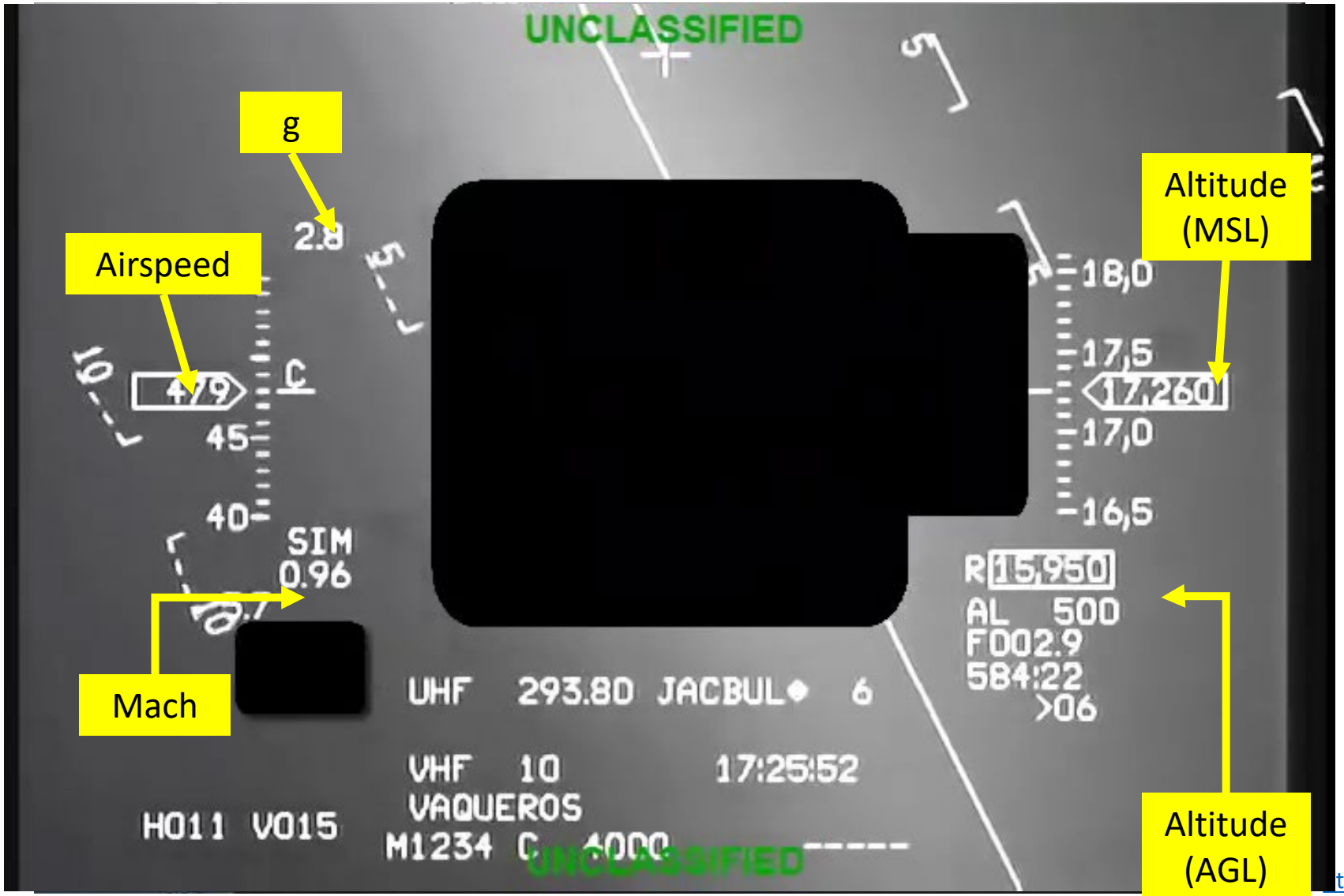
Source: https://www.ww2online.org

Mishap Report: F-16 Controlled Flight Into Terrain (CFIT)



https://youtu.be/WkZGL7RQBVw

Mishap Report: F-16 Controlled Flight Into Terrain (CFIT)



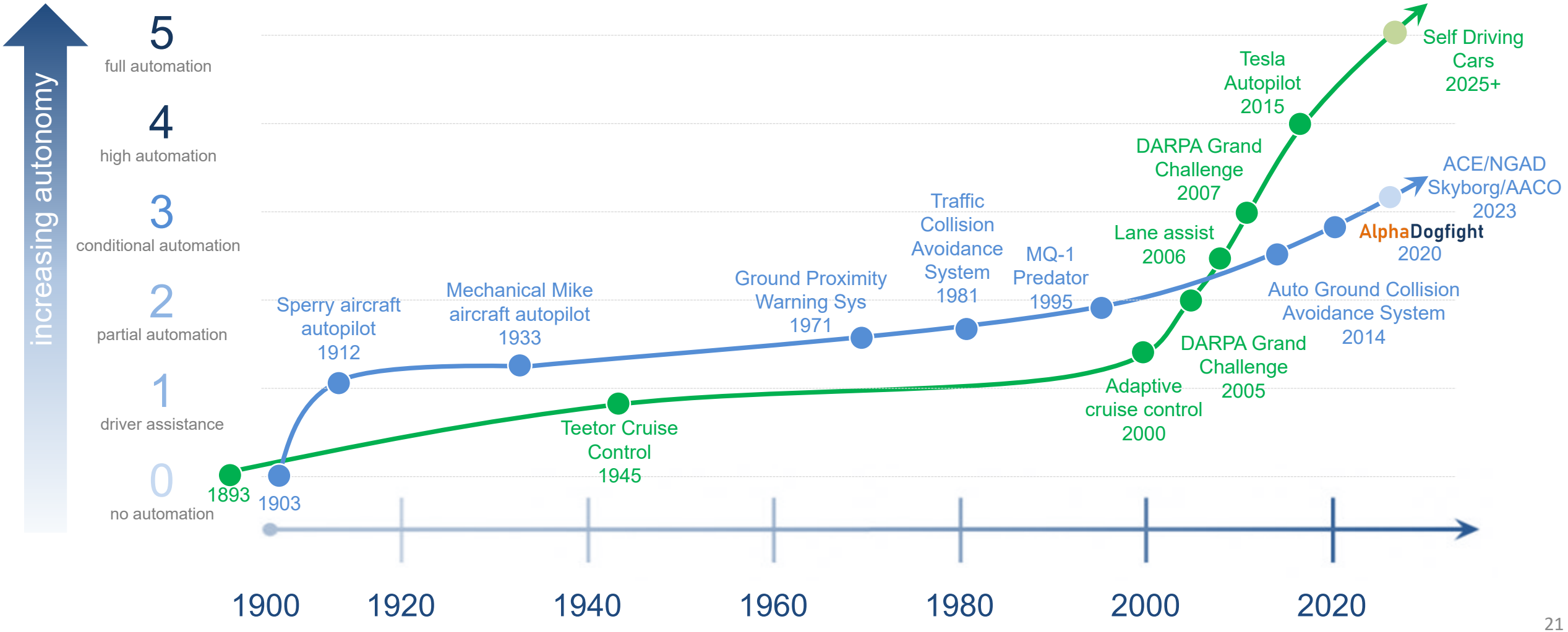
https://youtu.be/WkZGL7RQBVw

Source: U.S. Navy

Source: Scaled Composites

Autonomy was born in the air, now behind, why?

Commercial vehicle autonomy closing in on full autonomy [5]… ***…while aerial combat autonomy struggles to reach conditional autonomy [3]***



What is the vehicle autonomy approach to get to level 5?

A matter of quantity of quality data → Optimized for time to Market

- 2 challenges must be solved to get to market as fast as possible:
  - Performance: is it capable of handling what's required
  - Trust: expect it to deploy capability as required

by the way: these are the same characteristics we build/assess while training human pilots

Insight in their metric

What is measured is managed -- so what do they measure?
Assessment rolled up into a single metric:

Miles per Disengagement

miles

Source: Rand 2016

Table 1. Examples of Miles and Years Needed to Demonstrate Autonomous Vehicle Reliability

Statistical Question	Benchmark Failure Rate: (A) 1.09 fatalities per 100 million miles?	(B) 77 reported injuries per 100 million miles?	(C) 190 reported crashes per 100 million miles?
How many miles (years[a]) would autonomous vehicles have to be driven...			
(1) without failure to demonstrate with 95% confidence that their failure rate is at most...	275 million miles (12.5 years)	3.9 million miles (2 months)	1.6 million miles (1 month)
(2) to demonstrate with 95% confidence their failure rate to within 20% of the true rate of...	8.8 billion miles (400 years)	125 million miles (5.7 years)	51 million miles (2.3 years)
(3) to demonstrate with 95% confidence and 80% power that their failure rate is 20% better than the human driver failure rate of...	11 billion miles (500 years)		

[a] We assess the time it would take to compete the requisite miles with a fleet of 100 autonomous a day, 365 days a year, at an average speed of 25 miles per hour.

275 million miles (12.5 years)

- Waymo announced in January 2020 that its autonomous cars have driven **tens of billions** of miles through computer simulations and **20 million** miles on public roads in 25 cities"
- …and they are just 1 of 36 companies testing in California alone!

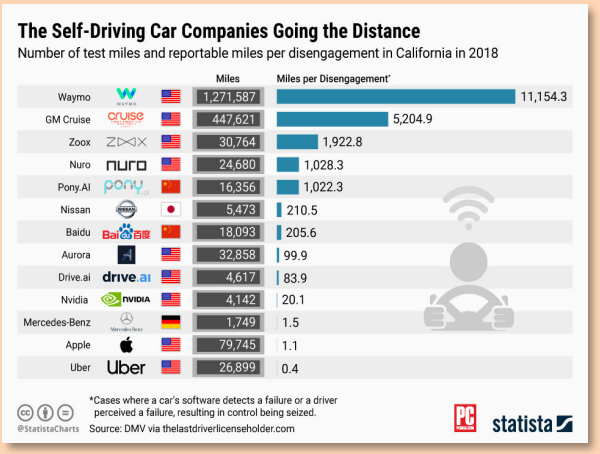
per disengagement

Also from RAND report: Miles alone not enough

Key findings

- Autonomous vehicles would have to be driven hundreds of millions of miles and sometimes hundreds of billions of miles to demonstrate their reliability in terms of fatalities and injuries.
- Under even aggressive testing assumptions, existing fleets would take tens and sometimes hundreds of years to drive these miles—an impossible proposition if the aim is to demonstrate their performance prior to releasing them on the roads for consumer use.
- Therefore, at least for fatalities and injuries, test-driving alone cannot provide sufficient evidence for demonstrating autonomous vehicle safety.

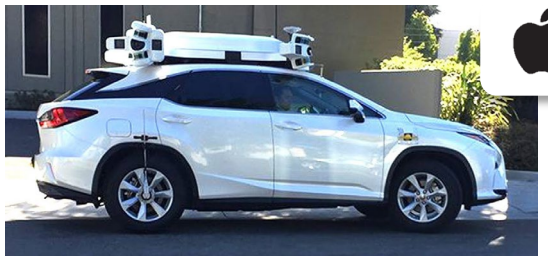
Need operator engagement to sufficiently build trust

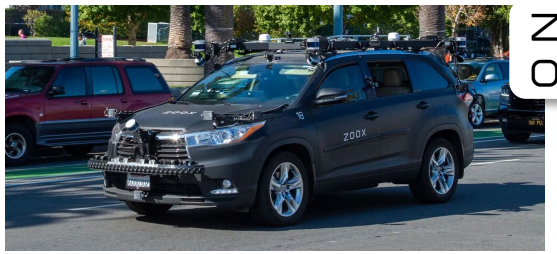


All the significant players are using a human on-the-loop configuration

Miles per disengagement verifies performance, builds trust, enables mileage volume

100s of autonomous vehicles testing on public roads in California alone…





…what does every single vehicle have in common?

human-on-the-loop

Another byproduct of approach is incremental feature rollout

Deploy minimum viable assistance autonomy features to build trust incrementally

LEVELS OF AUTONOMY

SCALE

0 No Automation
The full-time performance by the human driver of all aspects of the dynamic driving task, even when enhanced by warning or intervention systems *(i.e. Cars from 20 years ago)*

Human driver monitors the driving environment.

1 Driver Assistance
A driver assist system of *either* steering or acceleration/ deceleration using information about the driving environment and with the expectation that the human driver perform all remaining aspects the driving task.

2 Partial Automation
Driver assist systems are able to control both steering and acceleration/deceleration by using information about the driving environment while the human driver performs all remaining aspects of the driving task *(I.e. The system executes the steering/accerlation and deceleration, while the human monitors and acts as the fallback "system").*

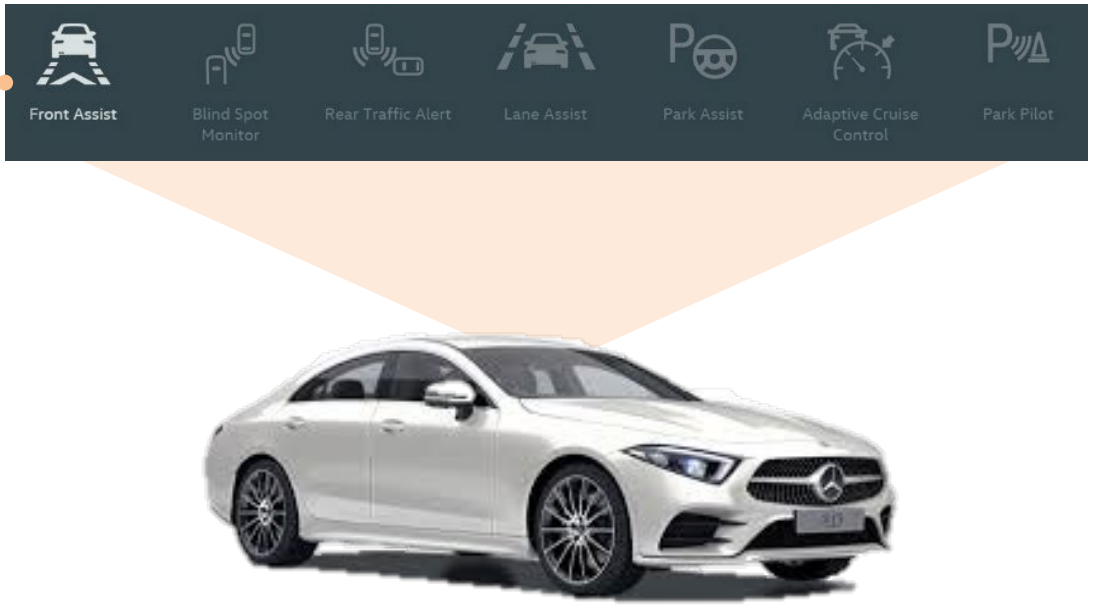
Driver is still the captain of the ship, but can relinquish certain controls at certain times to driver assist technologies.

3 Conditional Automation
automated driving systems perform all aspects of the driving task, with the expectation that the human driver will respond and intervene when needed.

4 High Automation
Performance by an automated driving system of all aspects of the driving task, even if a human driver does not respond appropriately to a system request or warning for human intervention. *(delivers the capability in a defined area/geofenced area and weather constraints)*

Automated driving system ("system") monitors the driving environment.

5 Full Automation
The full performance of driving by an automated driving system under all roadway and environmental conditions that can also be managed by a human driver, but human intervention is not needed. *(unrestricted area and weather)*



Incremental feature roll-out of increasing levels of automation: verifies performance while building trust

Super-human effectiveness of “pilot-assist features” demonstrated

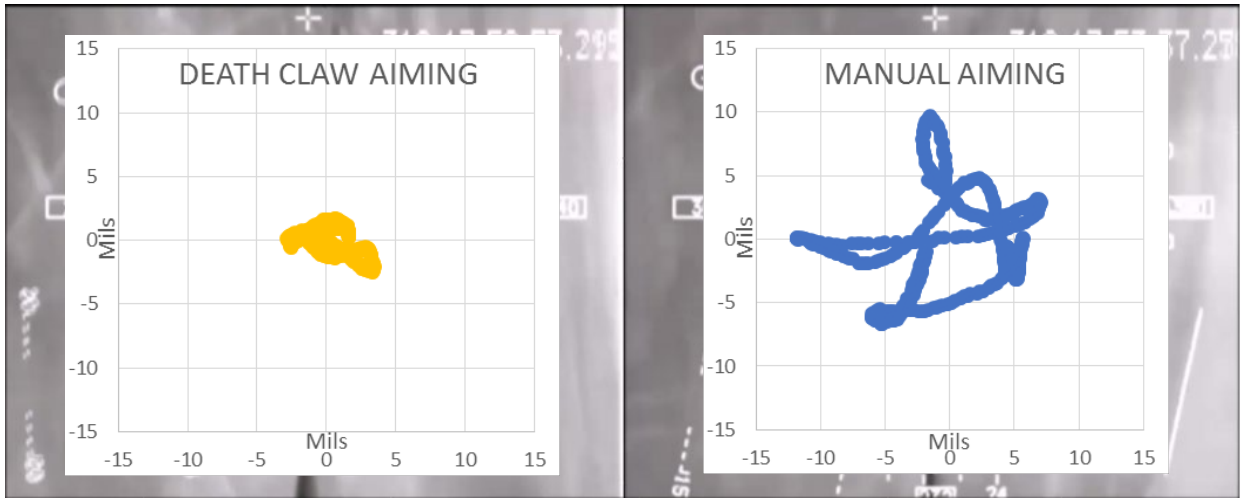
Death Claw and **AlphaDogfight** have both shown super-human performance in specialized tasks

ALIAS’s Death Claw

- Automated final portion of gun targeting to keep gun pipper on targets—a difficult & high workload task
- Flight tests using NF-16 VISTA test aircraft (USAF TPS) in November 2017
- Substantial performance improvements with unanimous, positive pilot comments

ACE’s AlphaDogfight

- Eight contractor teams competed for top AI agent before engagement with Weapons Instructor Course grad
- AI agent defeated Banger 5-0 in high aspect BFM due to superior aiming and decision making



Demonstrated effectiveness of automated aiming in air-to-ground and air-to-air employment

Air Force is currently betting on a different approach

Level 0-1 | Level 2-4 | Level 5

Commercial Self Driving Vehicles

To go from level 2 to level 5:

- ✓ Years: 20
- ✓ Vehicles: Thousands
- ✓ Miles: Hundreds of Millions

Why didn't commercial auto go straight to here?

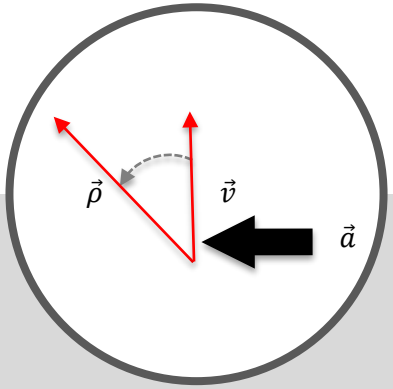
Air Combat Autonomy

Why are we going straight to here?

To go from level 2 to level 5:

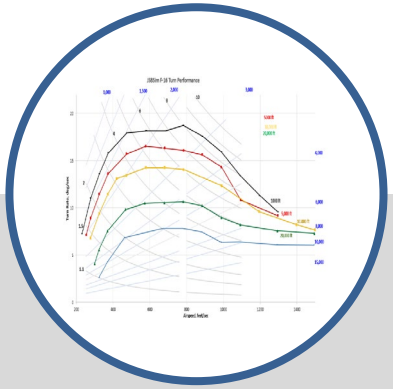
- ❑ Years: 0
- ❑ Aircraft: 1 (VISTA)…ACE will add 2xL-39
- ❑ Flight hours: Very few

The spectrum of AI approaches have already been explored



Rules-Based

- Traditional pro-nav autopilots
- Very quick to execute
- Low processing footprint
- Very predictable/explainable
- No novelty

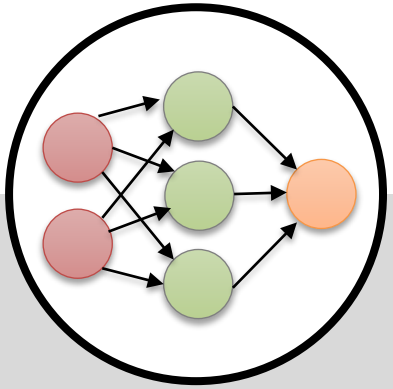


Expert System

- Rules-based GOFAI
- Quick to execute
- Low processing footprint
- Good explainability
- Limited novelty

Hybrid/Hierarchical AI

- Bootstrapped RL/Tree Search
- Abbreviated offline training
- Medium processing footprint
- Some explainability
- Potential for novelty



End-to-end AI

- Full Deep RL
- Long training timelines
- High processing footprint
- Very little explainability
- High potential for novelty

SOARTECH

Aurora FLIGHT SCIENCES
A Boeing Company

EpiSci

LOCKHEED MARTIN

Heron Systems Inc.

physicsAI

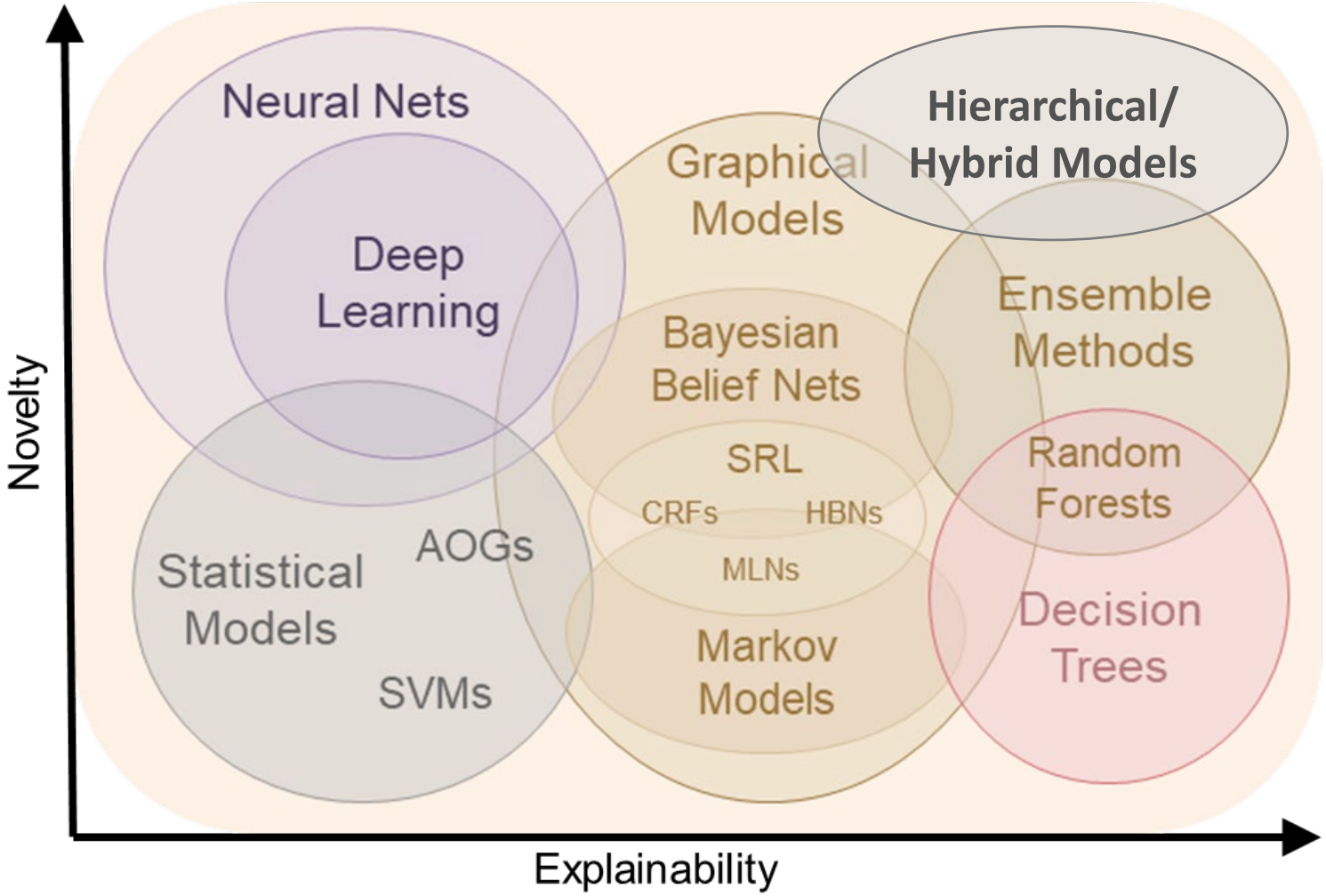
perspecta

AlphaDogfight Lessons Learned

Georgia Tech Research Institute

Primordial Labs

The tradeoff between novelty and explainability



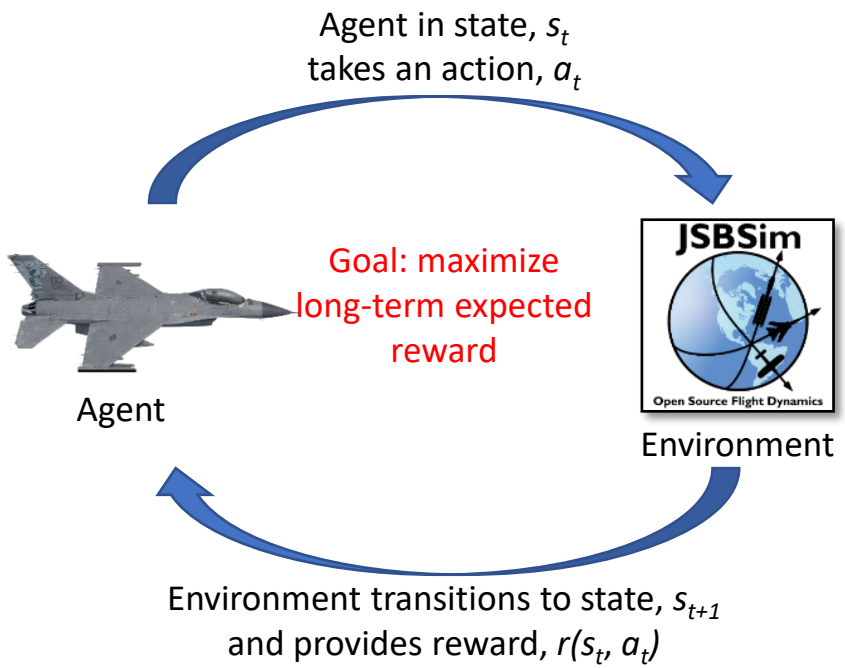
AOG: And-Or-Graphs
MLN: Markov Logic Network

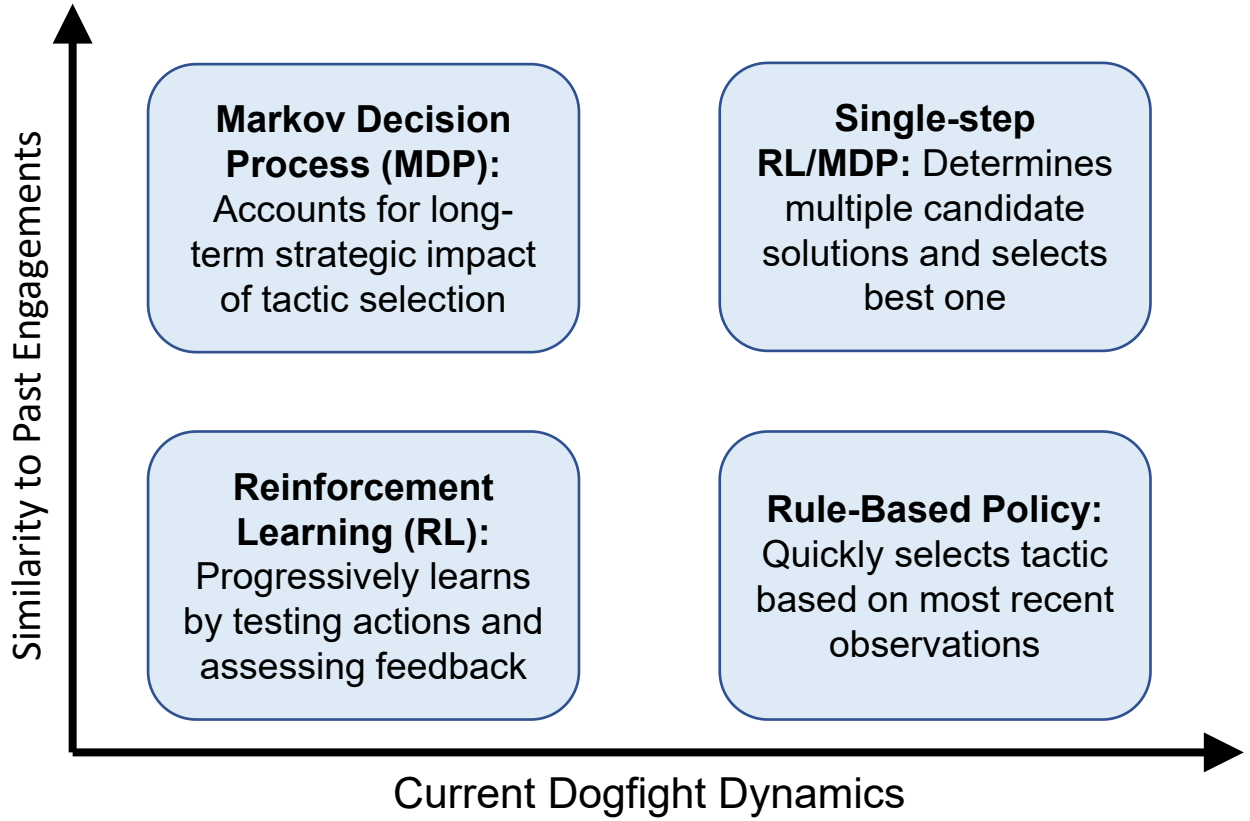
CRF: Conditional Random Fields
SRL: Statistical Relational Learning

HBN: Hierarchical Bayesian Networks
SVM: Support Vector Machine

Hierarchical/Hybridized solutions result in best of both worlds

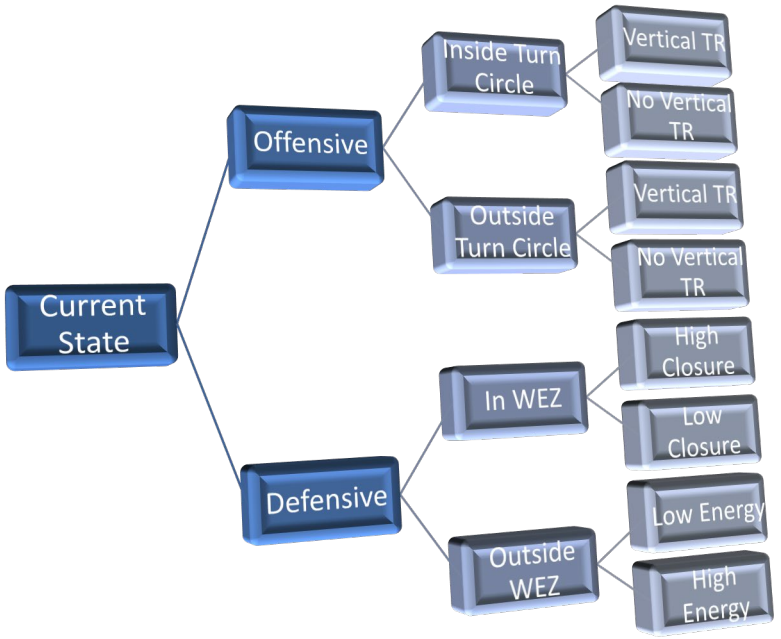
- A Hierarchical/Hybridized AI solution blends policies from:
  - Reinforcement learning artificial neural networks (referred to as multi-layered perceptrons)
  - Markov decision processes for the incorporation of uncertainties
  - Hierarchy and rules-based policies for explainability





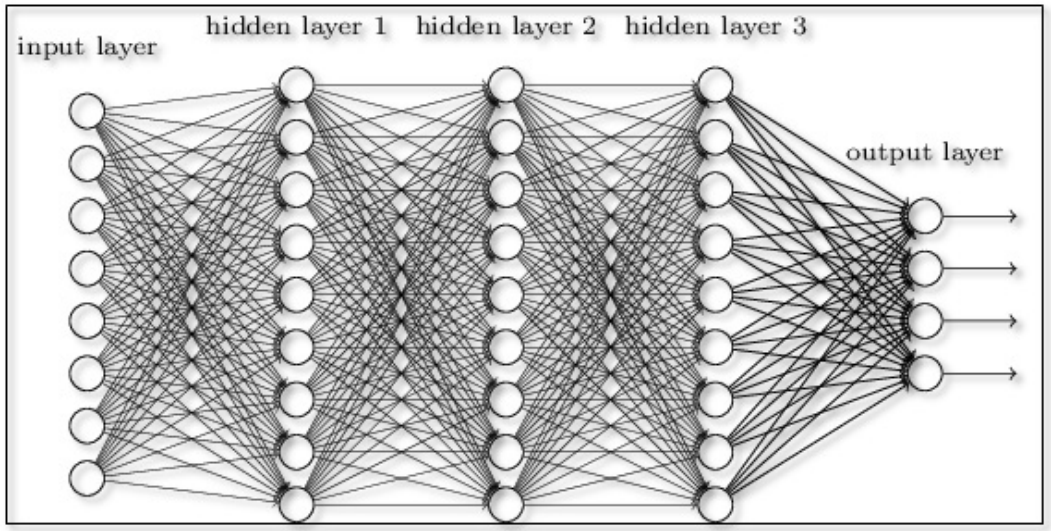
Hierarchical/Hybridized solutions combine two techniques

Expert System



- Performance increase via additional states
- Exponential design and test burden
- Straightforward verification; unbounded validation

Deep Reinforcement Learning

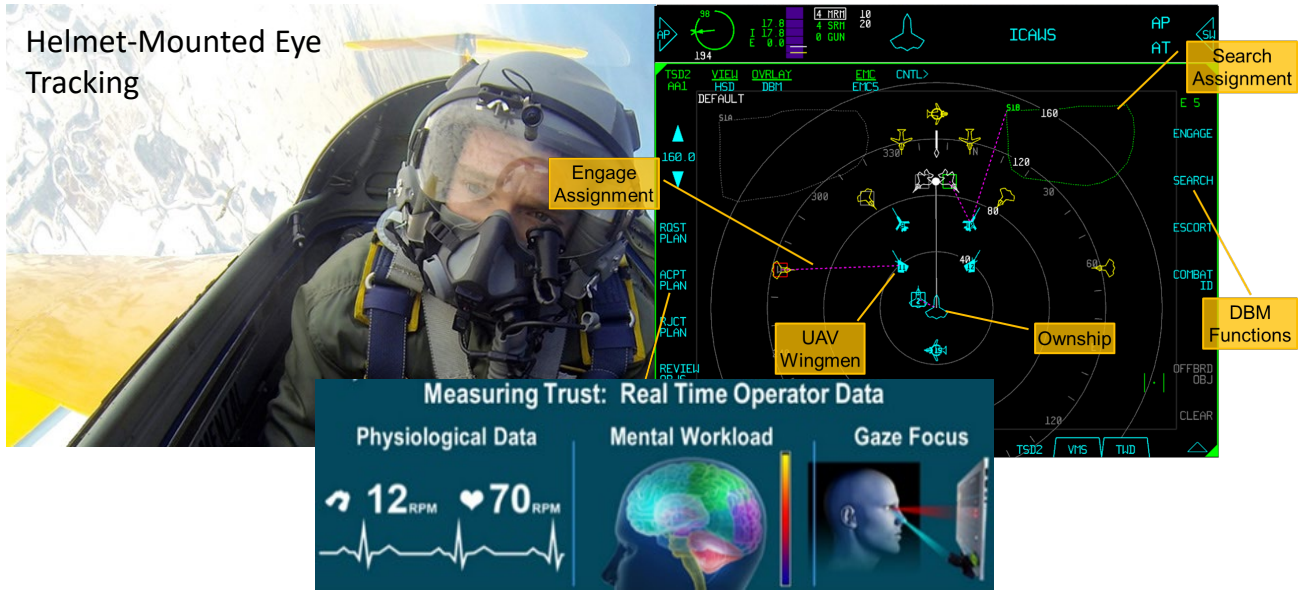


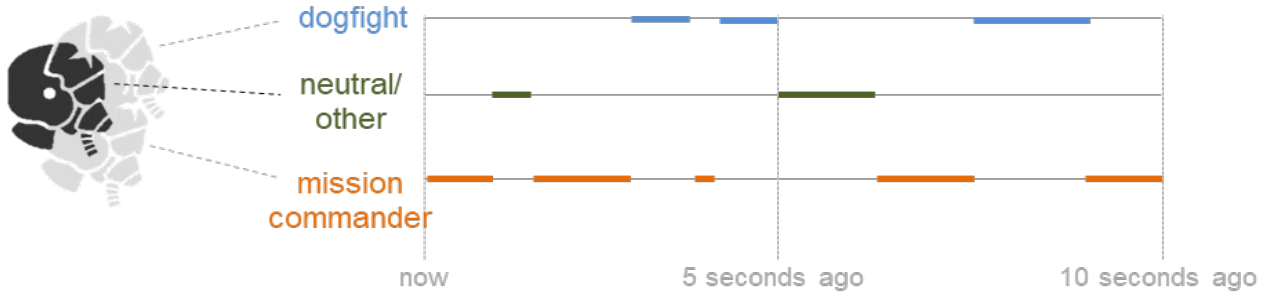
- Performance via simultaneous state processing
- Exponential training burden (simulation)
- Validation via training; unbounded verification

Must allow for, and even anticipate, simulation/reality mismatch

Reconsider Validation, Verification, and Accreditation (VV&A)

- One example of a trust methodology currently being explored





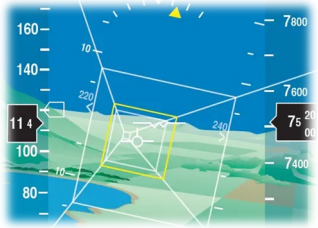
Divide the pilot's attention…

…measure workload distribution


AlphaDogfight Trials
VIRTUAL FINALS 8.18-20.2020
ARCNet
DARPA
APL
JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

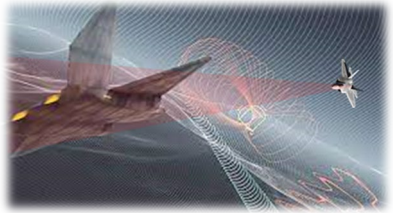
Autonomy applications for Air Combat go beyond dogfighting

Super-intelligent WVR dogfighting

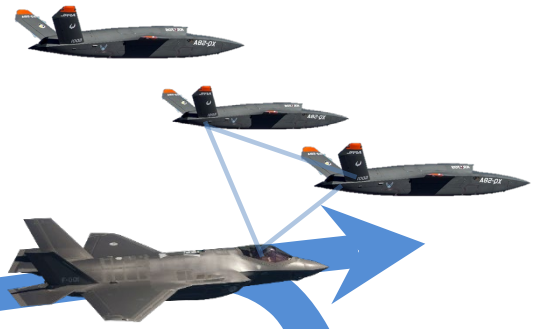


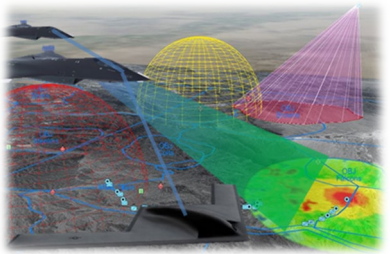
Flight Directors

Autonomous Threat Reactions



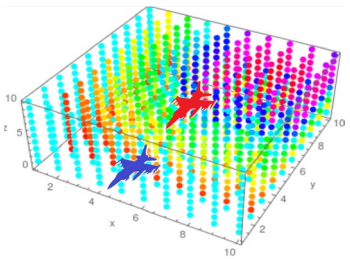
AI-enabled Electronic Warfare





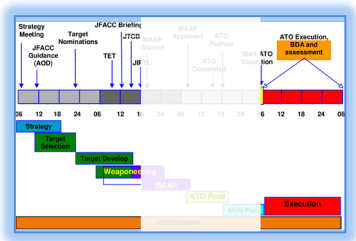
Human-On-The-Loop
Manned Unmanned Teaming
Experimentation

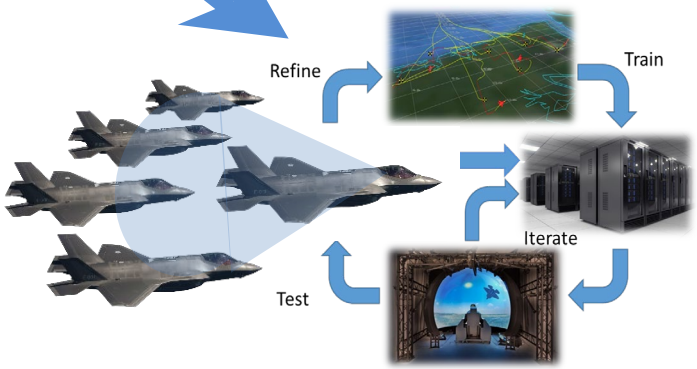
Briefing and debriefing



Signature Management

Air Tasking Order Generation





Fleet Learning

Our current systems are not as deterministic as we think

Consider performing envelope expansion testing on the following two aircraft

P-38 Lightning

- Federated subsystems
- Gaussian => Predictable
- Fewer test points
- Predict, Test, Validate Works
- Build-Up Approach Works
- Complicated

F-35 Lighting II

- Interconnected subsystems
- Non-Gaussian => Unpredictable
- More test points
- Predict, Test, Validate Tough
- Build-Up Approach Tricky
- Complex

Structure of a system: Simple

Simple | Complicated | Complex

Uncertainty

Single	**- Well understood - Quantitatively derived from first principles of physics**	**Single**	**None**

Structure of a system: Complicated

Simple | **Complicated** | Complex

Input → Function Model → Output

Uncertainty

Single Or Multiple	**- Well Understood** **- Quantitatively derived from:** • **First principles of physics** • **Physical certainties** • **Bounded assumptions**	**Single Or Multiple**	**Gaussian**

Structure of a system: Complex

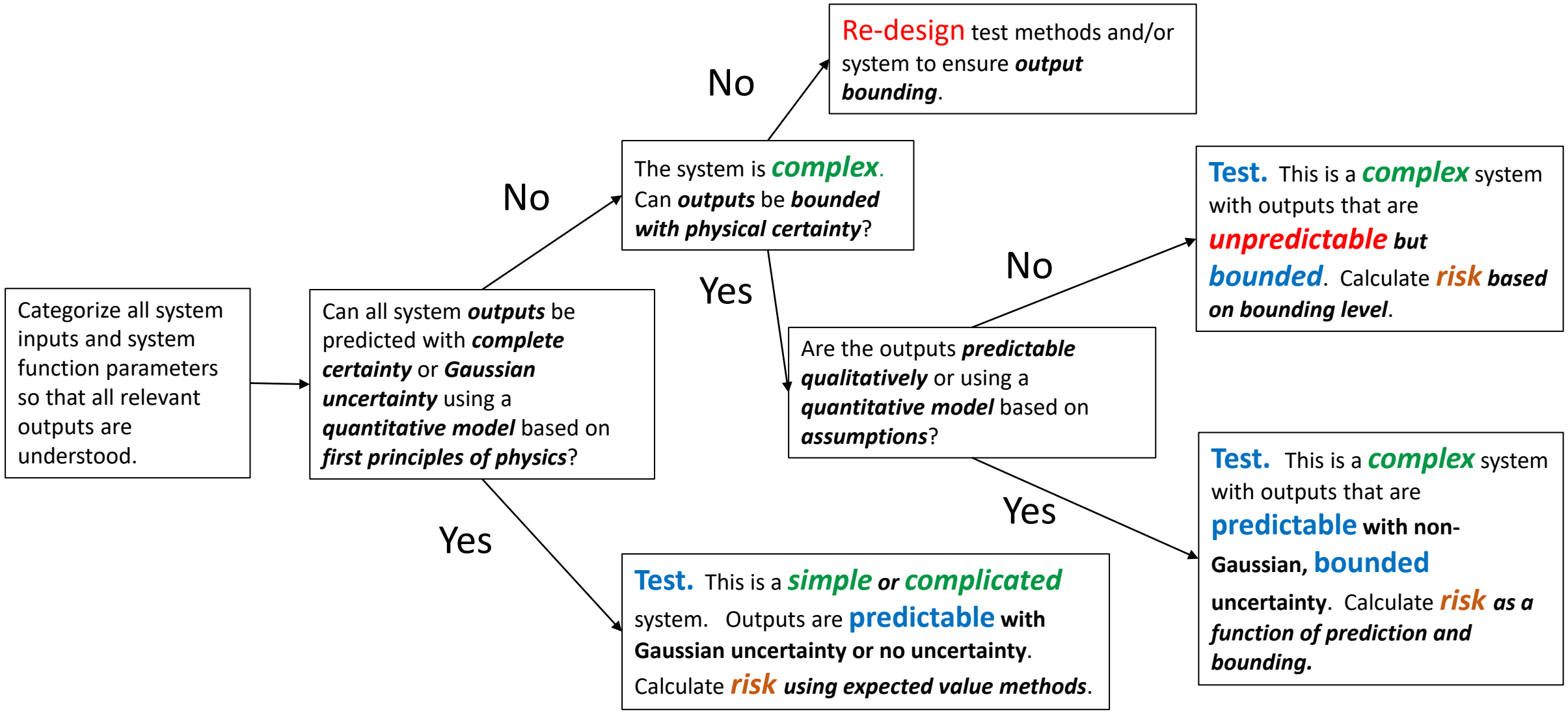
Simple Complicated Complex

Input → Function Model → Output

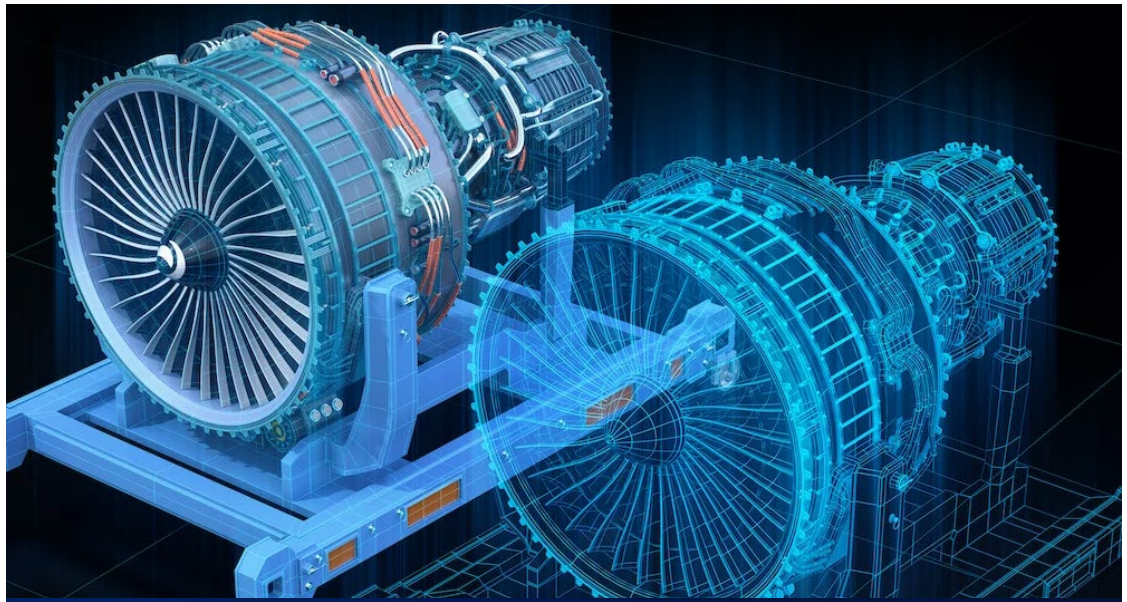
Uncertainty

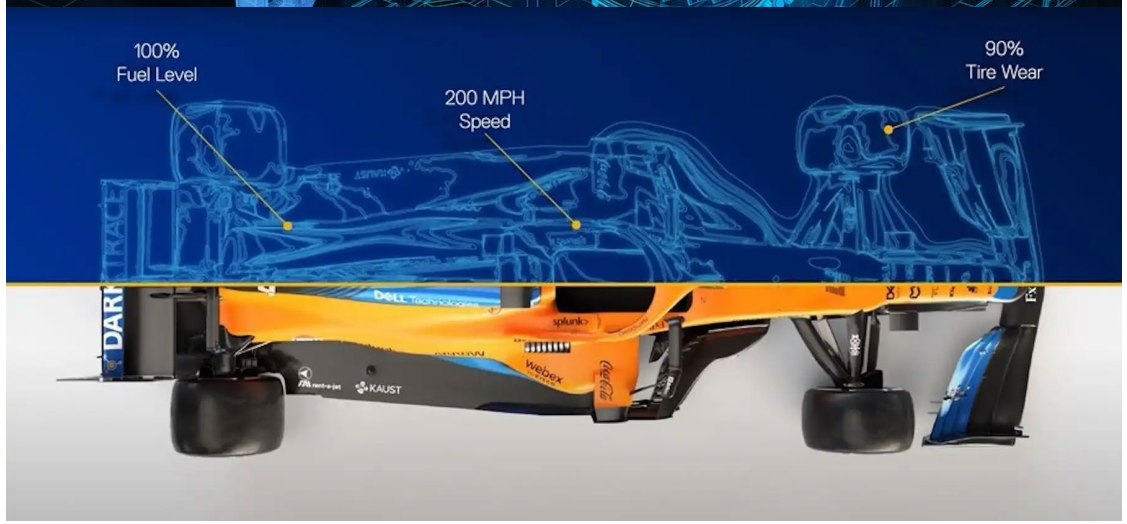
Single Or Multiple	**- No model** **- Qualitative model** **- Poorly understood quantitative model based on:** • **Assumptions** • **Not on first principles**	**Single Or Multiple**	**Non-Gaussian or Power Law**

Flight Test Suitability Flowchart



M&S, Digital Twins, and a continuous TEVVA Process





- Entry Criteria
- Steps
  1. Define the M&S/digital twin intended use
  2. Evaluate risk
  3. Define accreditation criteria
  4. Select the candidate M&S/digital twin
  5. Develop TEVV plans
  6. Perform TEVV
  7. Prepare accreditation support packages
  8. Accreditation assessment
  9. Accreditation decision
- Exit Criteria

Must convert all manned aircraft to autonomy test assets

Air Force does not have test assets to obtain flight hours required

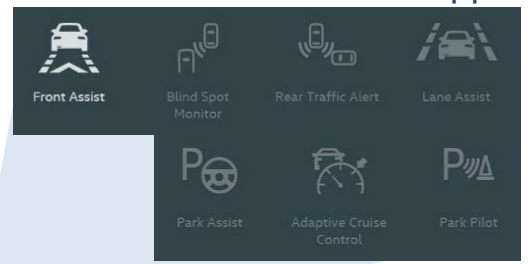
275 million miles required to prove a self driving vehicle is at least as trustworthy as a human – how many hours does the Air Force have?

Current test assets: 1? 2?

To go from level 2 to level 5:

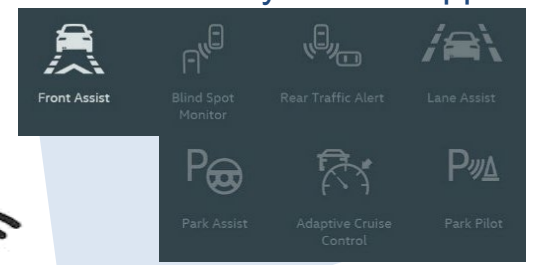
- ❑ Years: ??
- ❑ Aircraft: Full Squadrons
- ❑ Flight hours: 1000s
- ❑ Metric: Hours per disengagement

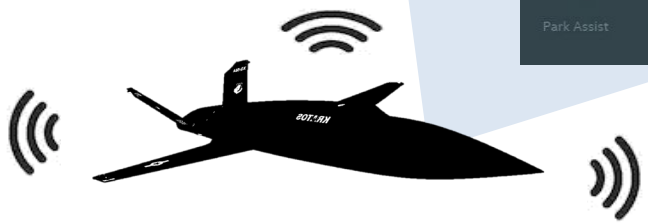
Pilot assistance combat apps



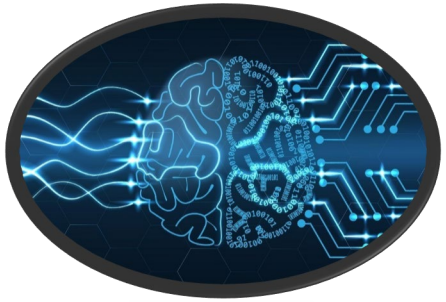
e.g. AutoGCAS, Death Claw, etc.

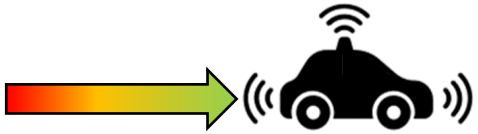
Full autonomy combat apps

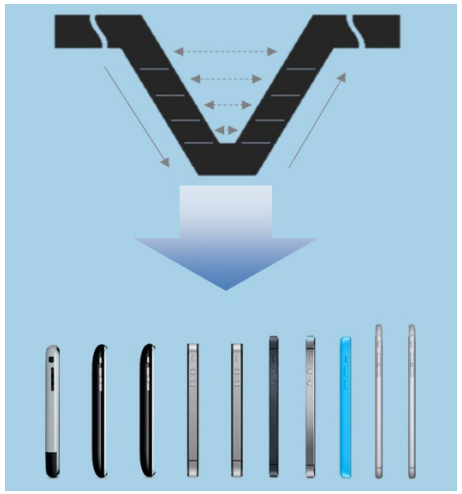




Greatly improved autonomy is coming to the cockpit







- The Augmented age of the 21st Century is upon us
  - Complete with AI
  - Hardware-enabled, software-defined systems
- Winning in autonomy is about volume: quantity of quality training data
  - Time to market: Must have a human-on-the-loop to adequately build performance and trust quickly
  - US military is taking a different approach – effectively skipping human-on-the-loop but this is changing
- We must dispel the myth that we cannot write a requirement for adaptability or perform continuous V&V
  - In the modern era full of complex systems we must modernize our Test, Evaluation, Validation, Verification, and Accreditation (TEVVA) processes
  - By choosing the right metrics, performing flight test suitability checks, and flipping the current role of M&S/digital twins we have a chance to tackle 21st Century challenges

Questions?