AI Assurance at RTRC

RTX

RTRC
Raytheon Technologies
Research Center

Brett Israelsen, Francesca Stramandinoli, Ganesh Sundaramoorthi

08/2023

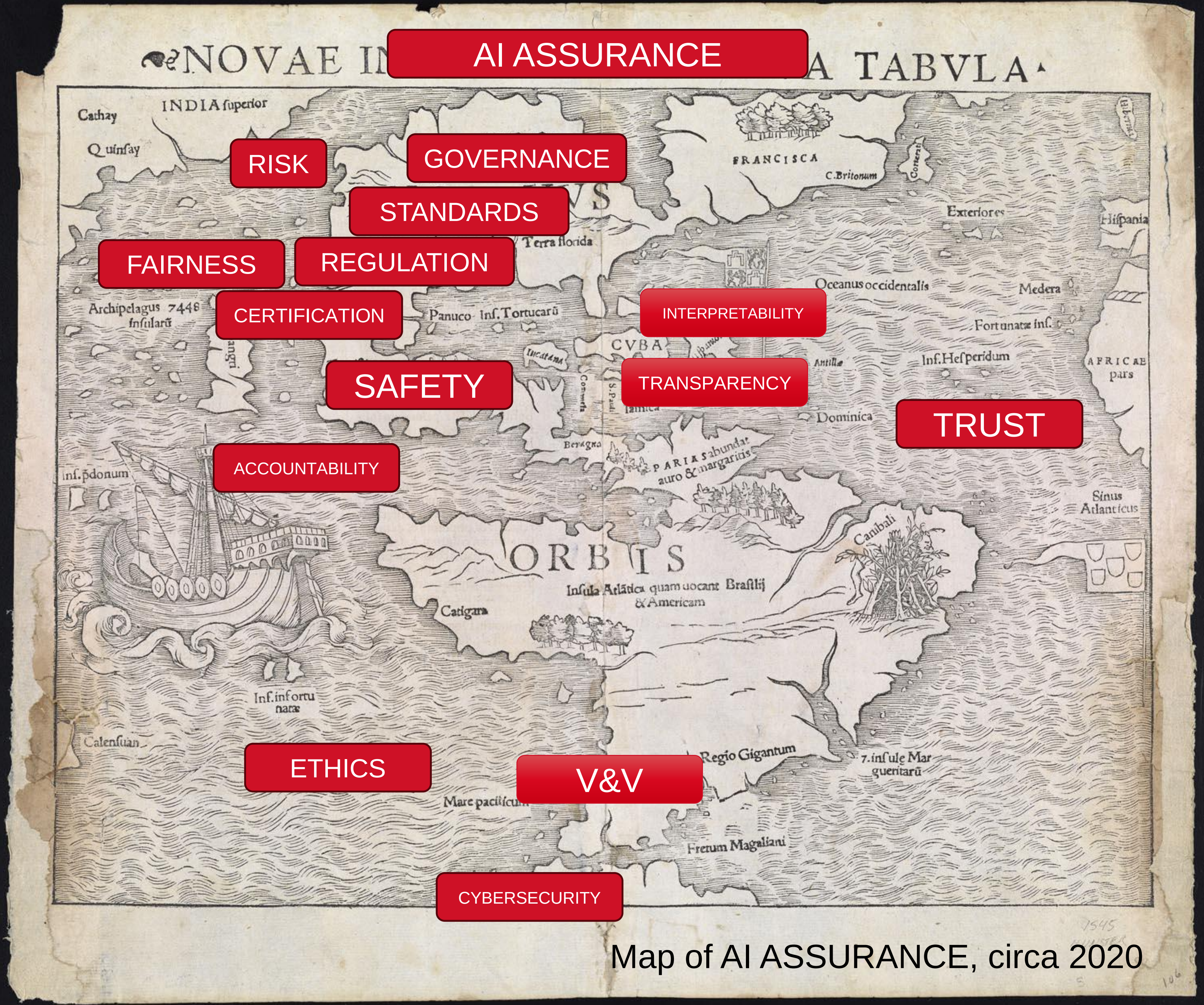
This presentation does not contain any export controlled technical data.

Trends influencing desire for more assurance

- Highly controlled environments → Complex uncontrolled environments
- Highly trained operators → Less-specialized operators
- Have to adapt to near-peer adversaries with similar technology
- Tasks delegated to systems are increasingly advanced

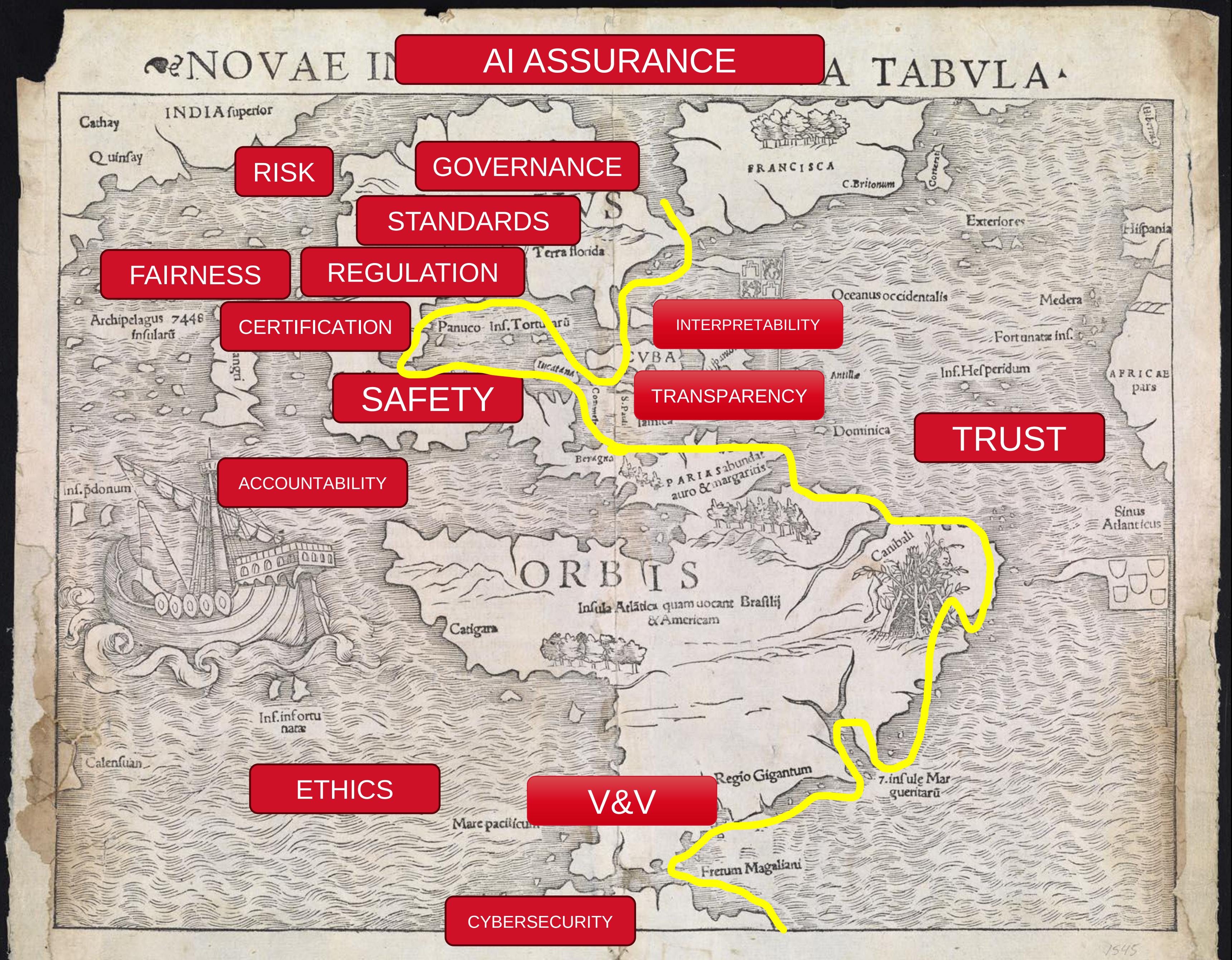
- Algorithms elude performance guarantees with current methods, **but** are required to address above points

Map of Americas, circa 1500



Map of AI ASSURANCE, circa 2020

- Our position is in many ways more complicated than map making
  - Concepts are not as concrete
  - Still trying to define what AI Assurance is
- We can certainly blaze our own trails, but:
  - Causes delays
  - Leads to oversight and errors
- Consensus Takes Time



Map of AI ASSURANCE, circa 2020

- Performed a trust-centered survey (Israelsen and Ahmed 2019)
- Identified agent-centered spectrum of assurances
- Useful for guiding R&D efforts, highlighting oversights/gaps
- There's still much more to discover

Trust vs. Trustworthiness

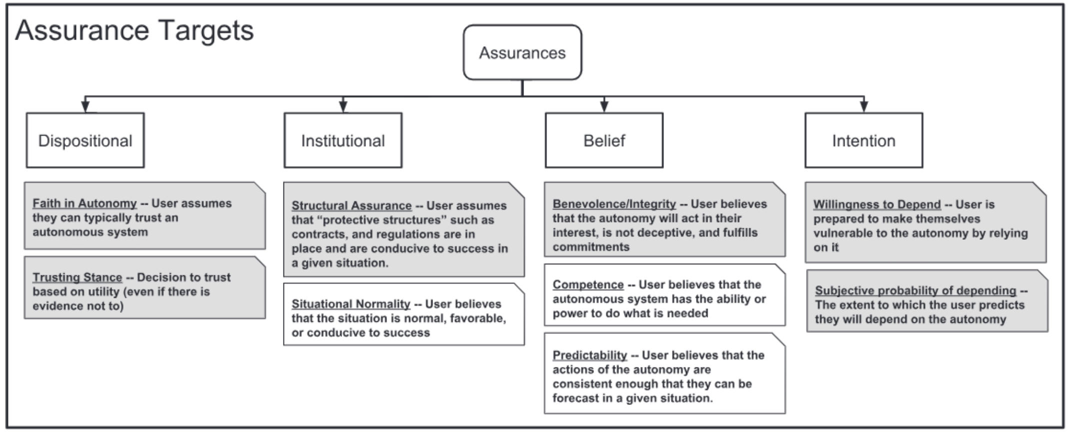
- Two distinct concepts
  - Trust
    - A psychological state in which an agent willingly and securely becomes vulnerable, or depends on, a trustee having taken into consideration the characteristics of the trustee
  - Trustworthiness
    - The degree to which an agent merits trust

- Addressing trust focuses on the user’s psychological state
- Addressing trustworthiness focuses on agent’s capability

Trust in AI

- We want to trust the AI/ML systems that we create
  - We require **assurances** to this end
- Interpersonal trust is a multi-dimensional construct
  - Human-AI trust is very similar
  - Dimensions include: competency, and predictability among others (McKnight 2001)
- Level of trust should be appropriate for:
  - A given **agent/system** (includes algorithms, data, and models)
  - In a given **context** (including things like environment and task)

Dimensions of Trust → Assurance Targets (Israelsen 2019)

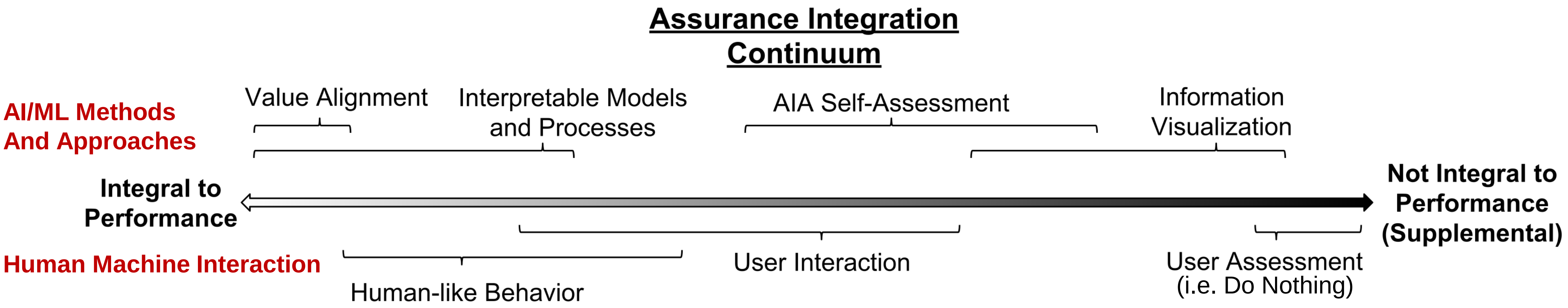


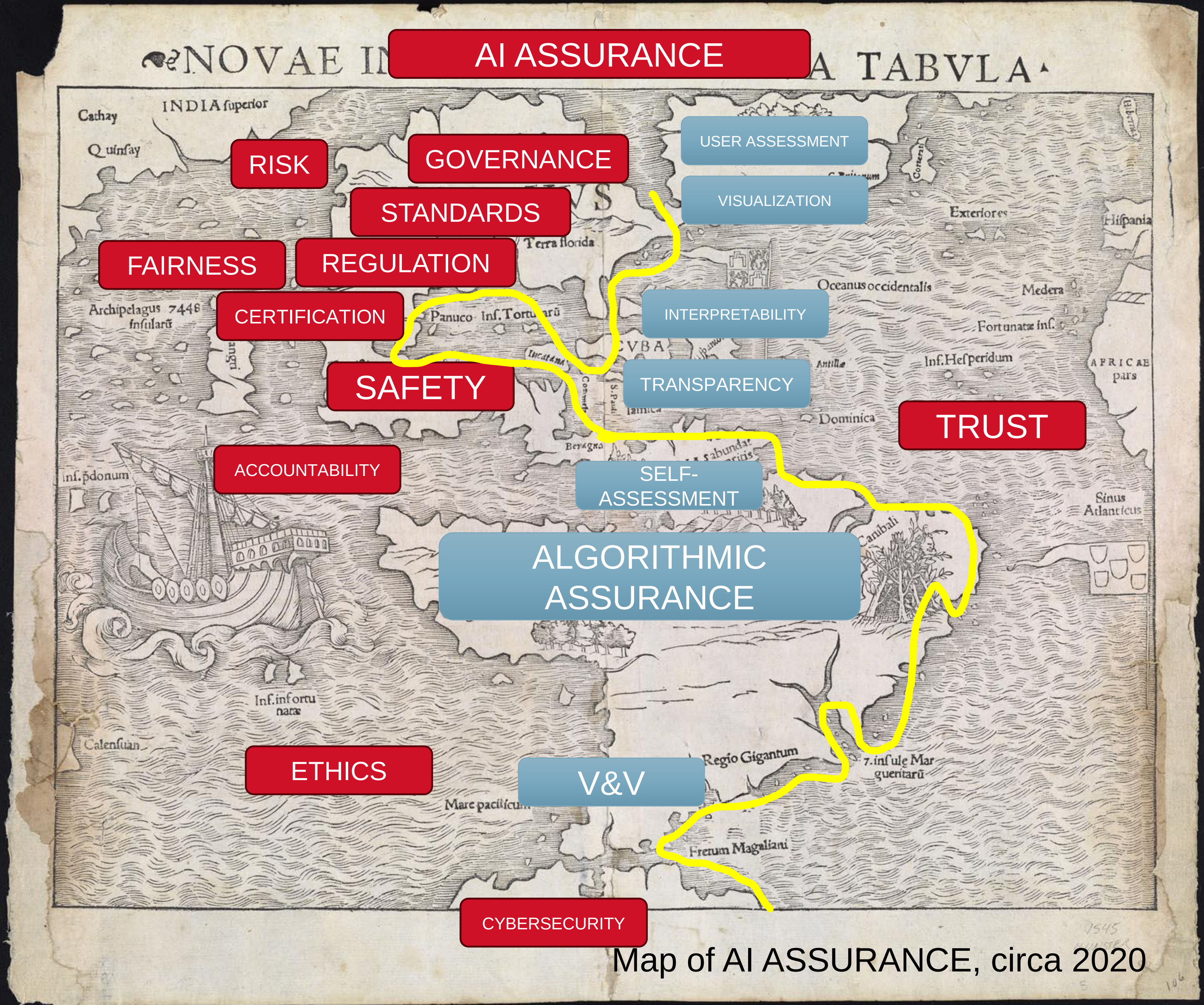
<u>Assurance can be thought of as evidence that trust is, in fact, merited and appropriate</u>
<u>However, trust is *not* the only factor</u>

[1]

Algorithmic Assurances (Israelsen 2019)

- Surveyed more than 200 papers
- An algorithmic assurance is an AI/ML agent/system property or behavior that can either increase or decrease user trust.
- Algorithmic Assurances can be applied at different levels of integration within an agent. These levels roughly encapsulate different technical approaches.



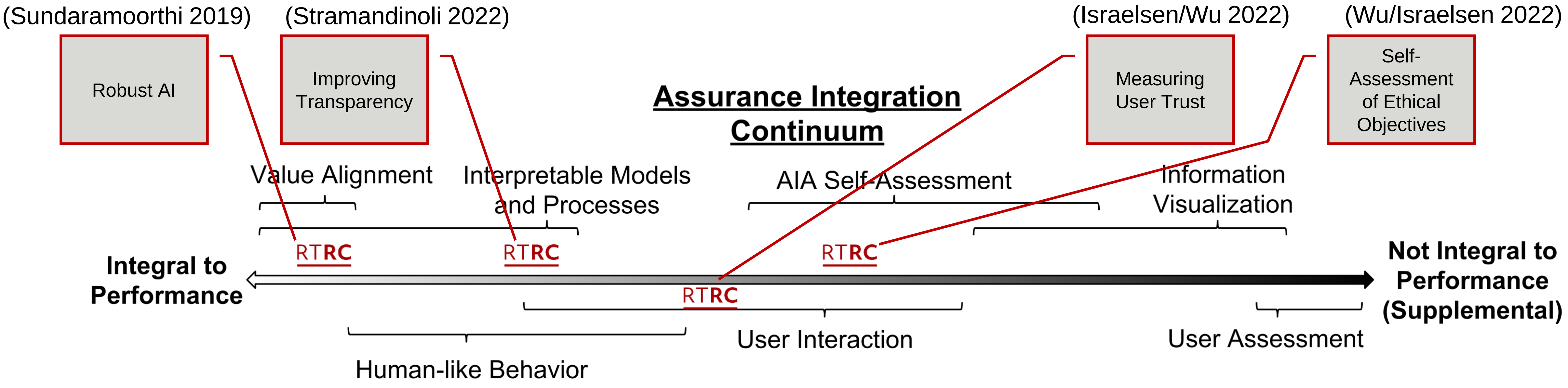


Map of AI ASSURANCE, circa 2020

- Performed a trust-centered survey (Israelsen 2019)
- Identified agent/system-centered spectrum of assurances
- Useful for guiding R&D efforts, highlighting oversights/gaps
- There's still much more to discover

RTRC Projects in the Assurance Landscape

- We are interested in methods/technologies across the spectrum below
  - We'll highlight a few today
- Most other talks today fall along this continuum as well



References

- D. H. McKnight and N. L. Chervany. 2001. What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology. *International Journal of Electronic Commerce* 6, 2 (2001), 35–59.
- Brett W. Israelsen and Nisar R. Ahmed. 2019. “Dave...I can assure you...that it’s going to be all right...” A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. *ACM Comput. Surv.* 51, 6 (January 2019), 1–37.
- Peggy Wu, Brett Israelsen, Kunal Srivastava, Hsin-Fu Wu, and Robert Grabowski. 2022. A Tiered Approach for Ethical AI Evaluation Metrics. Retrieved from https://www.researchgate.net/profile/Peggy-Wu-2/publication/358479807_A_Tiered_Approach_for_Ethical_AI_Evaluation_Metrics/links/6204312d075f695e892ea263/A-Tiered-Approach-for-Ethical-AI-Evaluation-Metrics.pdf
- Brett Israelsen, Peggy Wu, Katharine Woodruff, Gianna Avdic-McIntire, Andrew Radlbeck, Angus McLean, Patrick “dice” Highland, Thomas “mach” Schnell, and Daniel “animal” Javorsek. 2021. Introducing SMRTT: A Structural Equation Model of Multimodal Real-Time Trust. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI ’21 Companion), Association for Computing Machinery, New York, NY, USA, 126–130.
- Francesca Stramandinoli, Brett Israelsen, Peggy Wu, Kishore Reddy, Frank Tanner, Laura Strater 2022. User-intuitive Explanations for Increasing the Transparency of Autonomous Agents1st Annual Homeland Defense Awareness Symposium. https://media.defense.gov/2022/Jul/14/2003035169/-1/-1/0/HDAS%202022%20-%20STRAMANDINOLI%20%20-%20RTX%20HDSA%20SYMPOSIUM%20FULLPAPER%20V1%20FINAL.PDF
- Wang & Sundaramoorthi,Translation Insensitve CNNs, arXiv 1911.11238, 2019
- Khan et al., “Shape-Tailored Deep Nets,” arXiv 2102.08497, 2021

Francesca Stramandinoli

Increasing the Transparency of Autonomous Agents (ITAA)

RTRC Team: Brett Israelsen, Kishore Reddy, Francesca Stramandinoli, Peggy Wu
Raytheon Team: Laura Strater, Frank Tanner

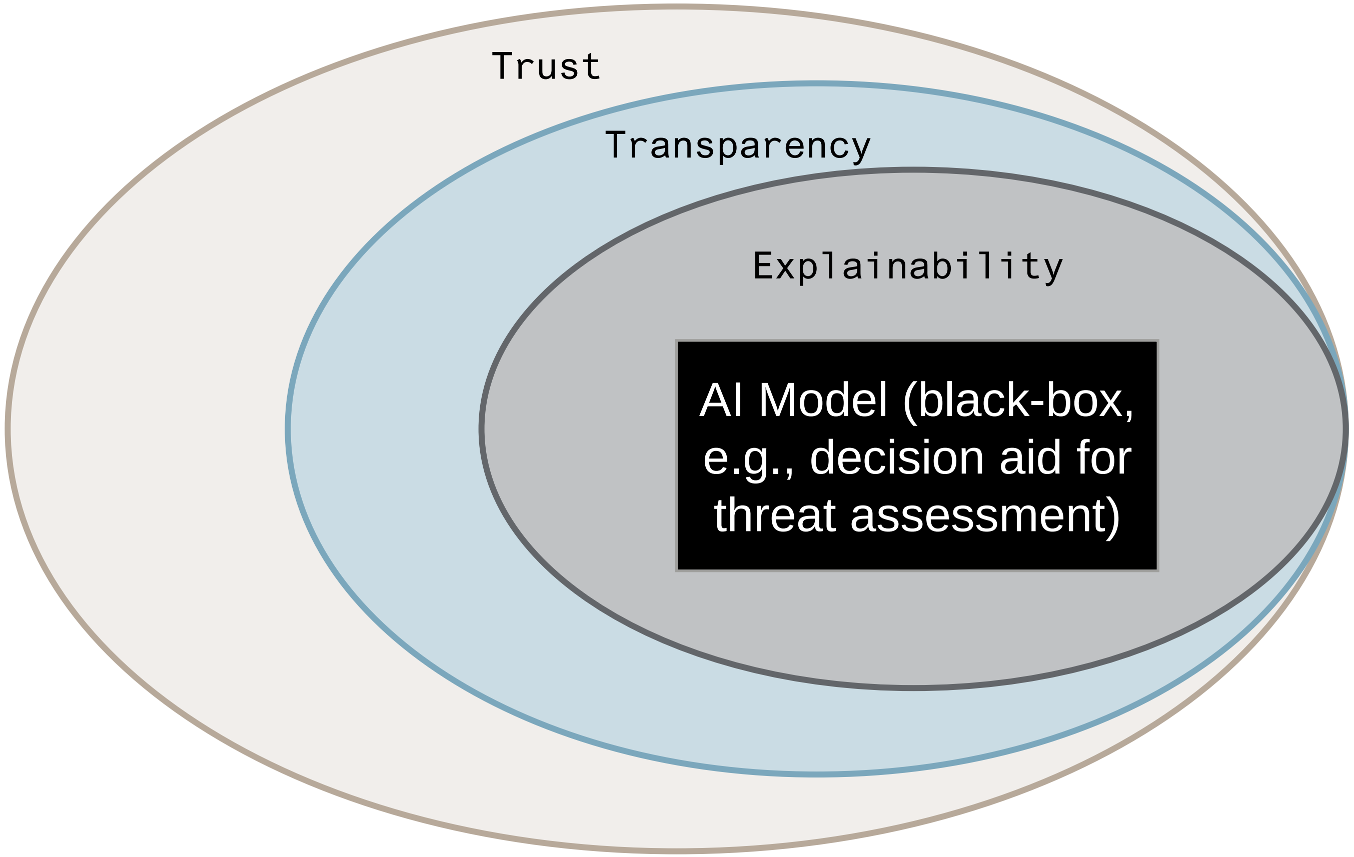
RTX RTRC

Explainability, Transparency, Trust

Explainability: Describe **WHY** a specific decision/recommendation is made.

Transparency: Does the explanation give the user a clear idea on **how the system works** (capabilities/limitations)?

Trust: Does the explanation **provide confidence** to the user in the recommender system?



Problem Statement

Increase Transparency of Autonomous Agents

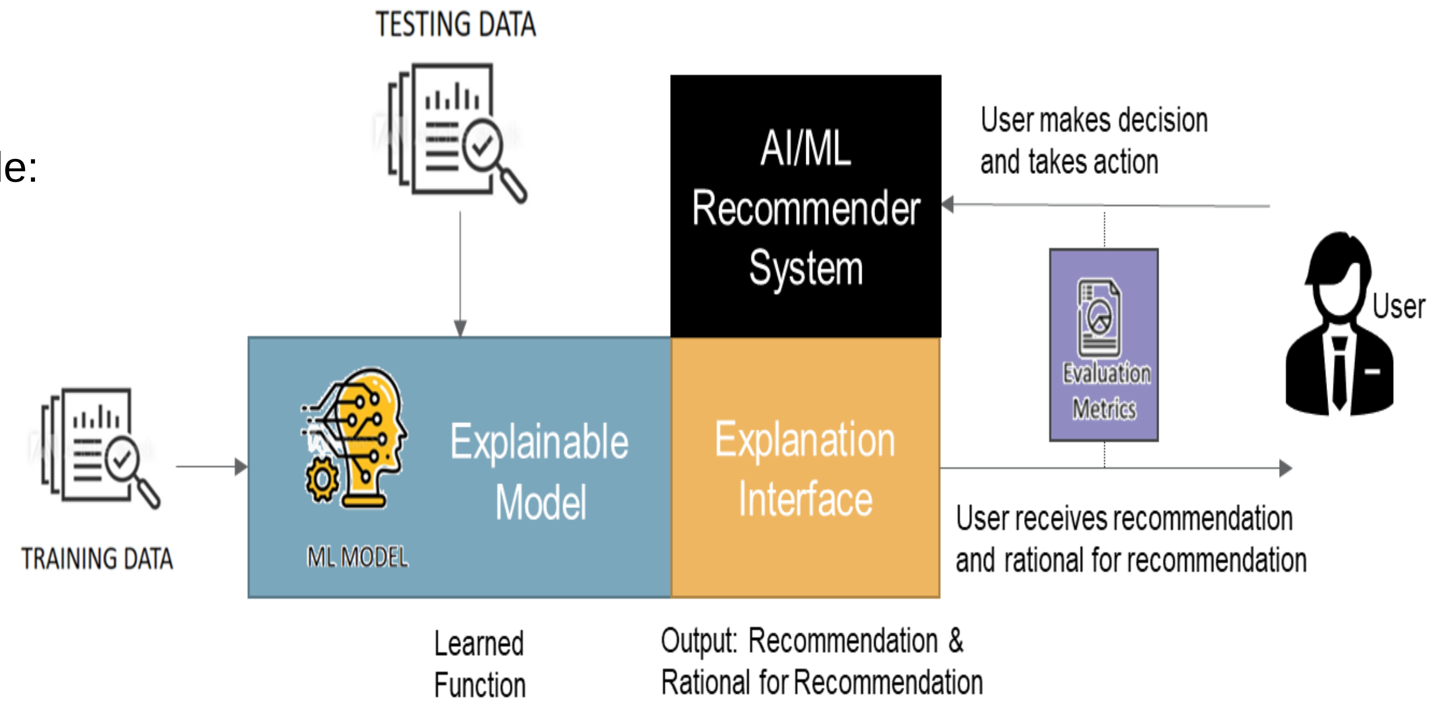
WHAT – Enable end users to determine when to trust the recommendation made by an AI / ML system and when to question it.

WHY – Improve Human + Autonomy decision cycle:

- **Efficiency** (faster decisions)
- **Effectiveness** (better decisions)

HOW – Leverage **Explainable AI (XAI)**:

- Models to generate explanations (AI/ML)
- Explanation Interfaces (HMI)
- Evaluation Metrics (AI/ML & HMI)

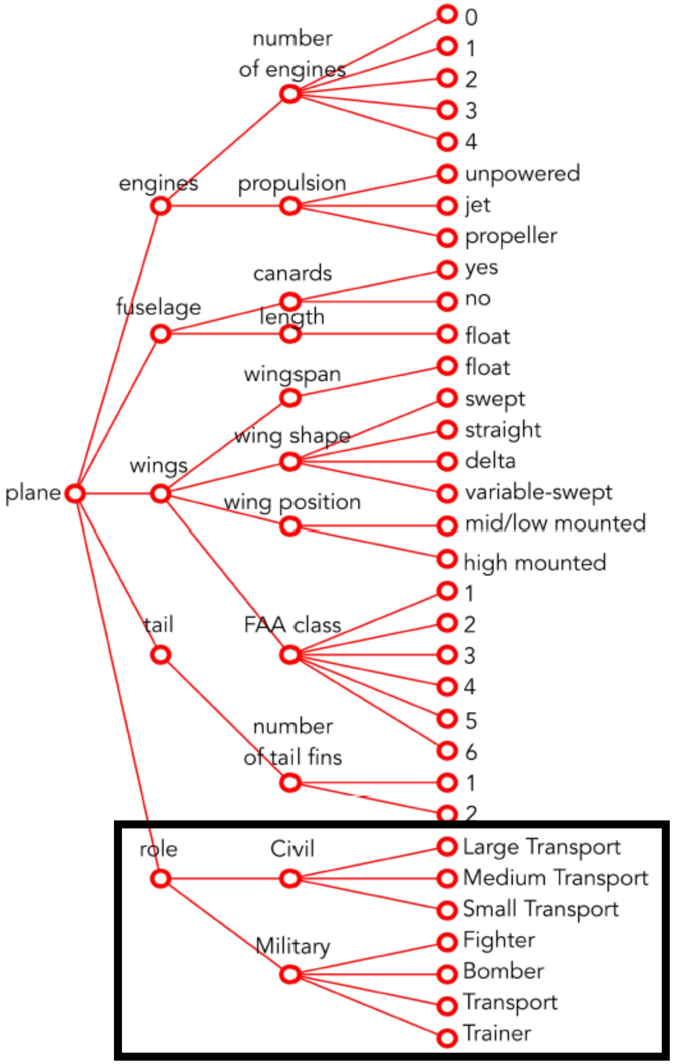


Motivating Use Case

Classification of Aircraft Role based on RarePlanes Data

- **Real data**:
  - **253 Maxar WorldView-3 satellite scenes** spanning 112 locations and 2,142 km^2 with **14,700 hand-annotated aircraft images**
- **Synthetic data**:
  - generated via AI.Reverie's simulation platform (based on unreal engine) and features 540,000 synthetic satellite images with **~630,000 aircraft annotations**
- **10 attributes** (from an overhead perspective)

Plane Attributes



Source: Shermeyer, J., Hossler, T., Van Etten, A., Hogan, D., Lewis, R. and Kim, D., 2021. Rareplanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 207-217).

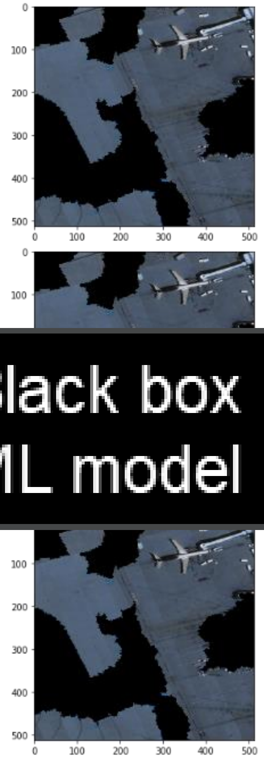
Feature-based Local Explainers

Local Interpretable Model-agnostic Explanations (LIME)

Input Image

Inpu

…the input image into …pixels: groups of …g pixels with similar …and brightness



Black box ML model

b) Create a collection of artificial photos by replacing random super-pixels of the original photo with black color

c) Use black-box model to predict classes of new images generated at step b)

Civil Aricraft / Military Aircraft

Black box ML mod… …ction

Prediction

d) Leverage input generated at step b) and output generated at step c) to train the explainer and generate the explanation

Explainer (e.g., LIME)

Explanation: image that represent the presence / absence of super-pixels that are most important to make a particular decision (featured-based explainer)

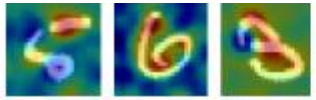
Need: focus more on the human side, aligning the generation of the explanation with the mental model of the final user.

User-intuitive Explanation Generation

Research Trends

- Future research in XAI will focus more on the human side, emphasizing the human-machine interactions and aligning the generation of the explanation with the cognitive model of the final user.
  - **Imagery data**: from **Saliency Maps** to **Concept Attribution**

Saliency Maps (SM): a map which highlights the contribution of each pixel at the prediction.

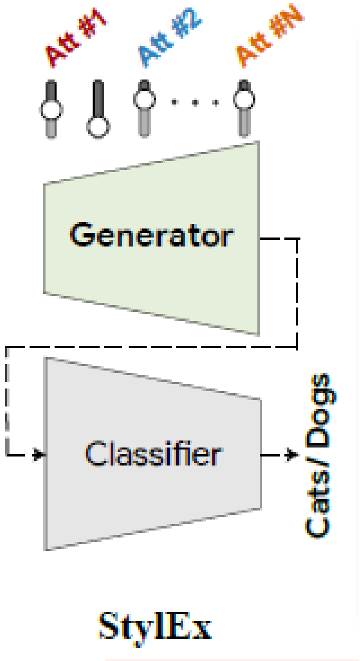


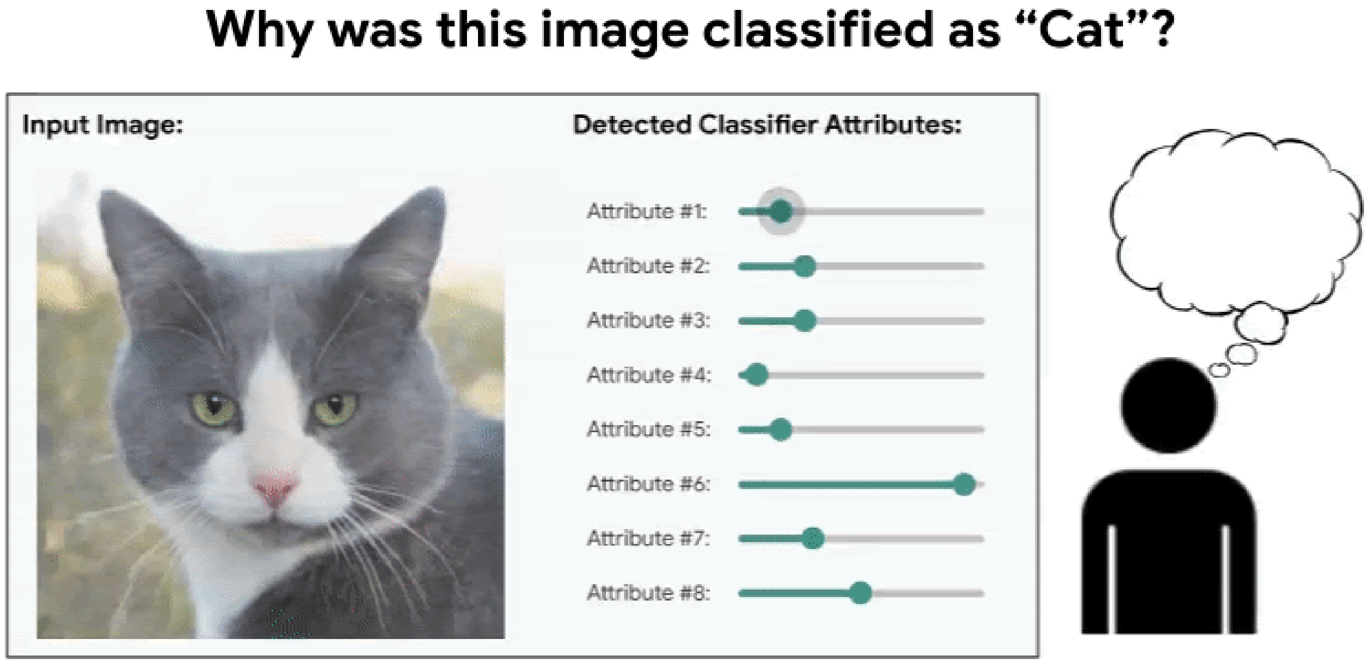
Concept Attribution (CA): quantify how much a concept example (e.g., stripes) has contributed to a particular class prediction (zebra).

Source: Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2021). Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*.

StylEx

Introduces a method for discovering classifier-related attributes and use them for counterfactual explanation generation (show how manipulating attributes affects the classifier prediction, ***Had the input x been ˜x then the classifier output would have been ˜y instead of y***).





Source: https://ai.googleblog.com/2022/01/introducing-stylex-new-approach-for.html

Drawbacks
There is no guarantee that the automatically discovered attributes will be human interpretable. Requires resources (i.e., a human) to label the automatically discovered attributes. Demonstrated on concepts relative to animals, foliage, faces, and retinal pictures.

Traditional Training Doctrine

Warfighters are trained to detect and classify vehicles using **fundamental building blocks:**

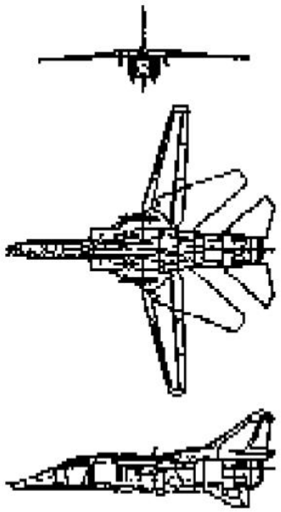
- Wings, Engine, Fuselage, Tail (WEFT doctrine)

WEFT FEATURES

		WINGS	ENGINES	FUSELAGE	TAIL
1.	Type	X	X		
2.	Position/Location	X	X		X
3.	Number	X	X		X
4.	Slant	X			X
5.	Shape	X	X	X	X
6.	Taper	X			
7.	Nose			X	
8.	Intakes		X		
9.	Rear			X	
10.	Exhausts		X		
11.	Mid			X	
12.	Cockpit			X	

TYPICAL AIRCRAFT DESCRIPTION FORMAT

MiG-27 FLOGGER D,J (MIKOYAN-GUREVICH)

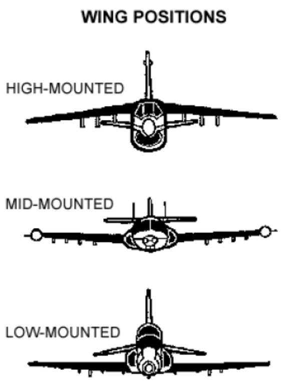


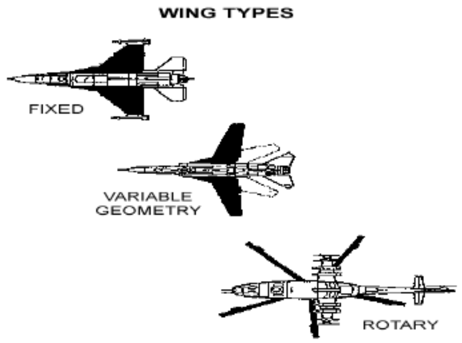
GENERAL DATA

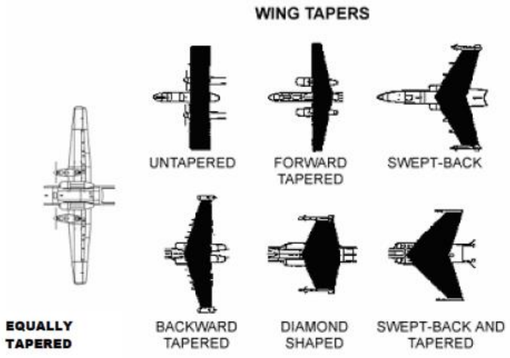
Country of Origin. CIS (formerly USSR).
Similiar Aircraft. MiG-23 Flogger B/E/G, F-111, Tornado, Su-24 Fencer.
Crew. One.
Role. Ground-attack, fighter.
Armament. Missiles, bombs, rockets, cannons.
Dimension. Length: 55 ft (16.6m).
Span: 46 ft, 9 in (14.26 m).

WEFT DESCRIPTION

Wings. High-mounted, variable, swept-back, and tapered with blunt tips.
Engine(s). One inside the body. Rectangular box-like air intakes forward of the wing roots. Single exhaust.
Fuselage. Long and tubular, except where air intakes give a box-like appearance. Long, downward-sloping, sharply pointed nose. Stepped canopy. Large, swept-back, and tapered belly fin under the rear section.
Tail. Swept-back and tapered tail fin with curved dorsal in leading edge and angular tip. Swept-back and tapered flats high-mounted on the fuselage with angular tips.





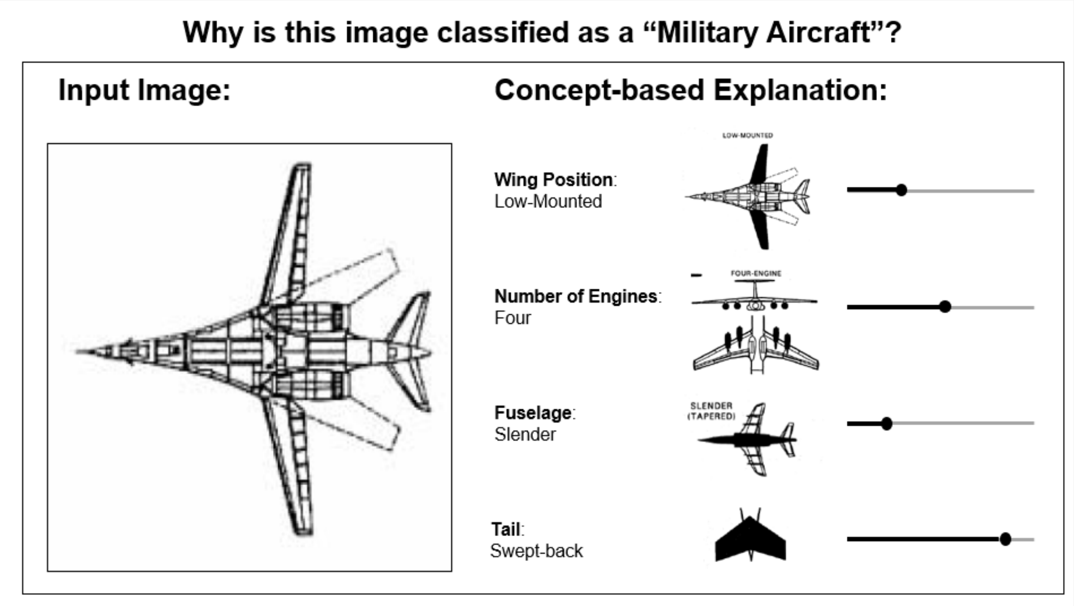


Leverage WEFT concepts/attributes for designing explainability methods aligned with user' internal representation of the problem.

Increasing Transparency of Autonomous Agents

Raytheon Technologies

<u>Problem:</u> Lack of interpretability leads to opaque decision-making systems that can negatively impact humans' trust in autonomous agents.



<u>Focus of project:</u>

1. **Develop** algorithms for user-intuitive explanation generation
2. **Define** multi-modal explanation framework
3. **Demonstrate** PoC validation of user-intuitive explanations

The video shows how the number of engines changes (from 2 to 4) on the plane.

Discriminator	Competitive Benefits
User-intuitive Explanation Generation	• *Clear and easy-to-understand explanations can reduce cognitive workload of human operators for validating decisions made by AI / ML models.*
Multi-modal Explanation Generation	• *Produces coherent explainable decisions combining reasons from individual AI / ML models. This enables to improve the transparency of AI / ML models, and therefore, improve the effectiveness and efficiency of human + autonomy decision cycle.*

CRAD Prospects: *Advance the state-of-art in trusted AI. This gives the opportunity to generate materials for engagement with external customers.*
On-going pursuit: Pre-marketing: US Air Force Academy (USAFA); ATRWG and the National System for Geospatial Intelligence (NSG); ARL; AFRL.

Leverage WEFT concepts/attributes for designing explainability methods aligned with user' internal representation of the problem.

Ganesh Sundaramoorthi
Sr. Technical Fellow, RTRC

Robust AI

RTX RTRC

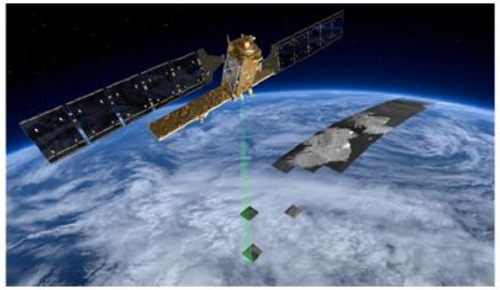
Vision: Assuring AI for Aerospace & Defense

Address Problems Limiting Deep Learning in Aerospace & Defense

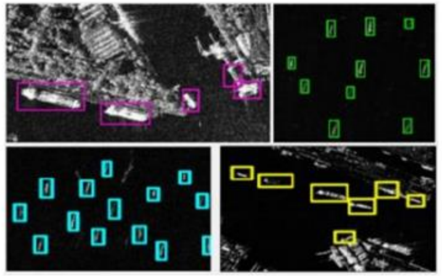
Challenges, Limitations of Existing Art

Sample Use-Case: ATR in Maritime

SAR Imaging



Detection

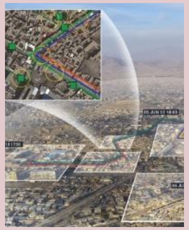


- Lack of robustness to image nuisances (viewpoint, illumination, occlusion, noise) and adversarial examples
- Lack of generalization and need for large datasets
- Heuristic design of deep learning (DL): no assurance
- Lack of explainability
- Expensive: Labor & Compute
- Not suitable for edge: Large size, weight and power
- Verification & Validation not possible yet

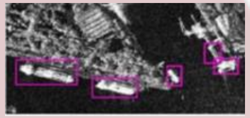
RTX

Auto Land / Cargo Loading

Automated Inspection



ISR (Situational awareness)



ATR

DARPA

IARPA
BE THE FUTURE

Office of Naval Research
ONR
Science & Technology

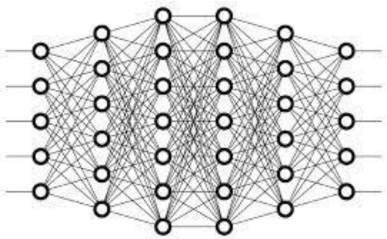
From DARPA GoL Program: *"Deep Learning practice outpaces theory, creating barriers to adoption in DoD.* GoL *seeks to develop theoretical tools that could advance existing DARPA AI programs."*

New theoretical tools for deep learning needed for Aerospace & Defense

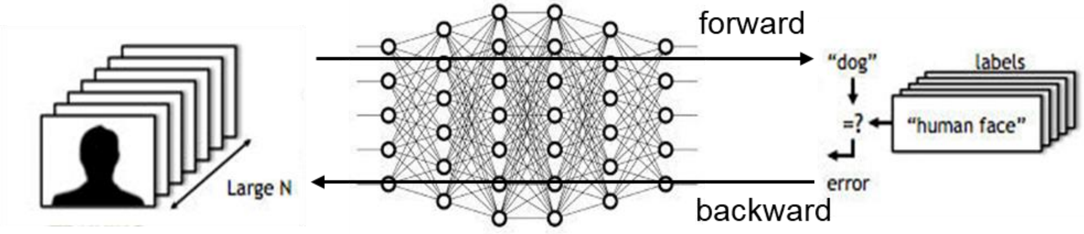
RTX RTRC

Three Components of Assuring AI

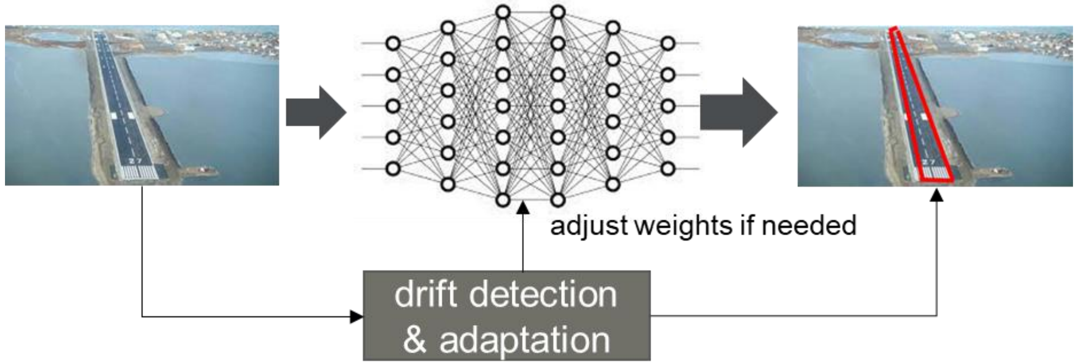
Architecture Design



Training Optimization Design



Online Network Adaptation

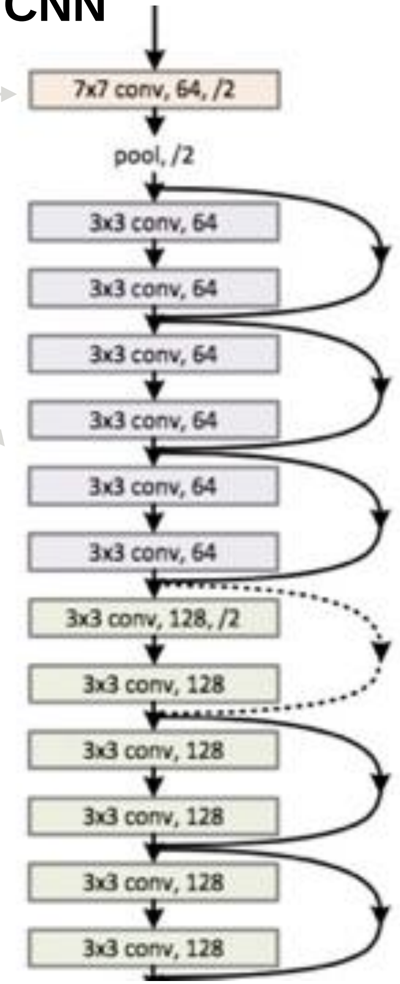


Assurance Through Architecture Design

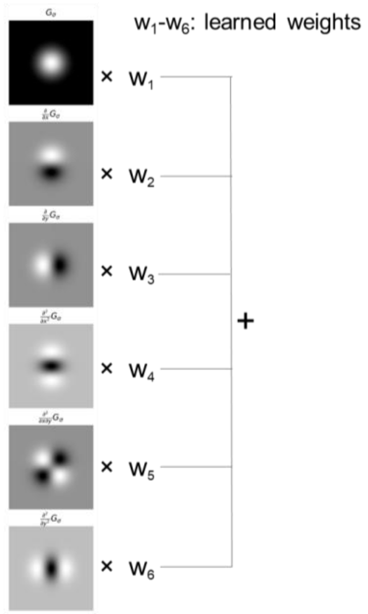
Lack of robustness is due in part to non-smooth kernels

Nuisance-Robust CNN

Replace conventional kernels with *Gauss-Hermite kernels*



Gauss-Hermite kernel



- New architecture with theoretical performance guarantees: analytically provable ***robustness to a wide range of image nuisances***
- Reduces training requirements: in-built invariances

Empirical Validation on Benchmark Dataset (CIFAR-10):

Model	Accuracy	Sensitivity – Delta1	Sensitivity – Delta2
ResNet (conventional net)	88.22%	8.47%	21.74%
NR-CNN (robust net, ours)	**91.54%**	**2.86%**	**8.28%**

Wang & Sundaramoorthi,Translation Insensitve CNNs, arXiv 1911.11238, 2019
Khan et al., "Shape-Tailored Deep Nets," arXiv 2102.08497, 2021

NR-CNN Naturally Induces Robustness to Wide Range of Nuisances

Assurance Through Optimization Design

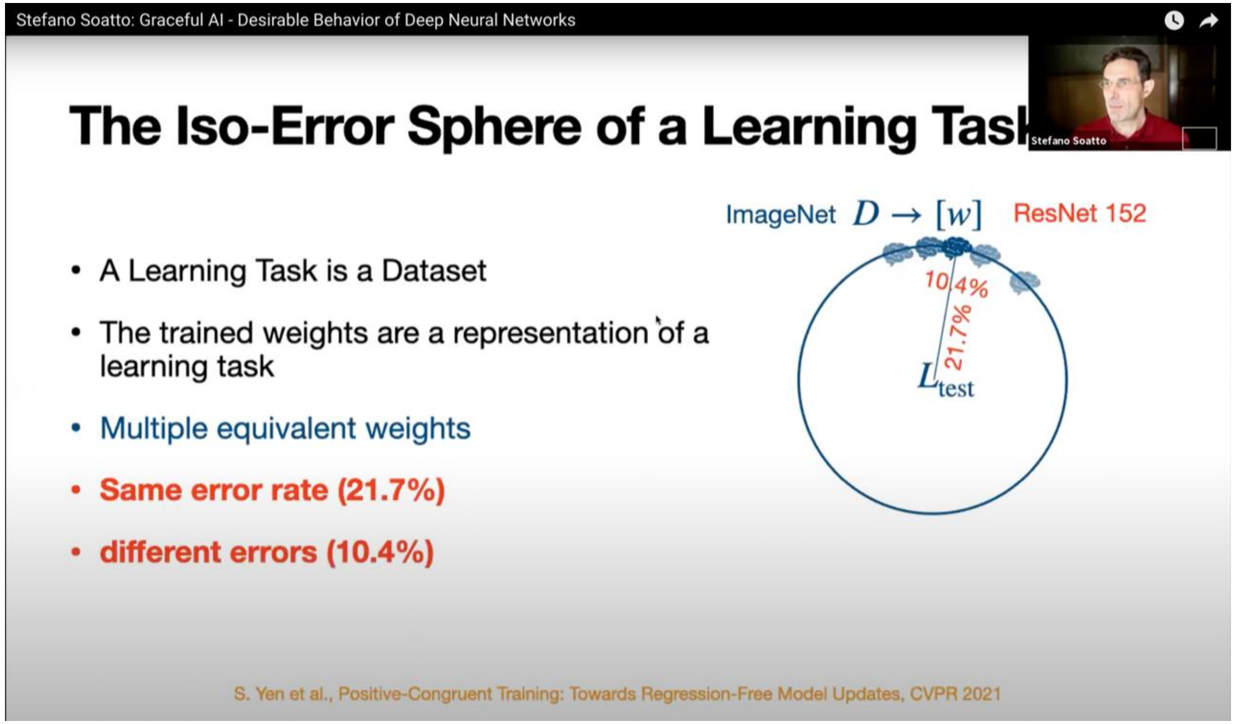
Variance of SGD: Well-Known

$$\theta_{t+1} = \theta_t - \eta g_t$$

Trial Runs of SGD on ImageNet / ResNet 152

Trial #	Error-Rate
1	21.70
2	21.72
3	21.74
4	21.71
5	21.73
6	21.73
Std dev	0.01

Why is this variance of concern? VP Amazon says:
- Models with nearly same accuracy disagree significantly
- Model updates can change seeds – resulting in disagreements
- Amazon customers lose trust in model



Optimization Variances Can Lead to Trust / Assurance Issues

Assurance Through Optimization Design

New Discovery: Unexpected Variance in SGD

SGD: $\theta_{t+1} = \theta_t - \eta g_t$

Perturbed SGD: $\theta_{t+1} = \theta_t - (\eta/k) \times (k g_t)$

(k is an odd integer)

Relative variance of gradient perturbations:

SGD	26.72
Perturbed SGD	2^{-23}

(Expt on CIFAR-10 / ResNet 50)

Accuracy variance from SGD

SEED	1	2	3	4	5	6	STD
$k = 1$	93.36	93.40	93.10	93.14	93.34	93.33	0.11
$k = 3$	93.49	93.37	93.08	93.68	93.16	93.12	0.22
$k = 5$	93.64	93.22	93.39	93.17	93.26	93.42	0.16
$k = 7$	93.36	93.31	93.12	93.23	93.14	93.28	0.09
$k = 9$	93.87	93.55	93.08	93.35	93.42	93.41	0.24
$k = 11$	92.99	93.31	93.49	93.48	93.14	93.56	0.21
STD	0.27	0.10	0.16	0.19	0.10	0.13	

Accuracy variance from Perturbed SGD

Deep Net Optimization is Not Stable – Theoretical Analysis; see our papers:

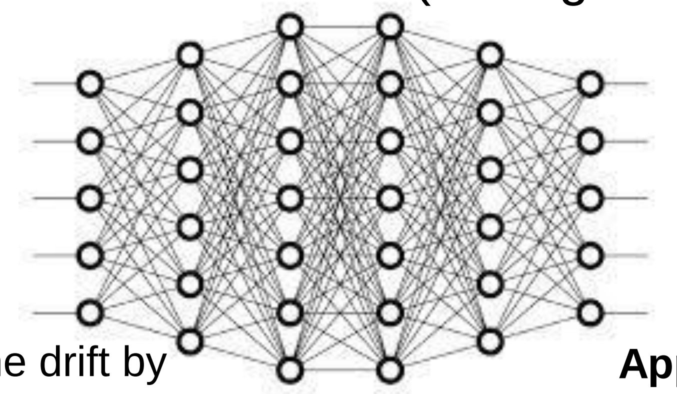
- Y. Sun et al., "Surprising Instabilities in Training Deep Networks and a Theoretical Analysis," NeurIPS 2022
- Y. Sun et al., "A PDE Explanation of Extreme Instabilities and Edge of Stability in Neural Nets," JMLR, 2023 (under revision)

Stability of Deep Net Optimization Is Important for Assured AI

Assurance Through Online Adaptation

Part of the way to adaptation: drift detection

Data Drift Detection (During Inference)



Runway (90% confidence)

Approach 1: Determine drift by comparing image to training data

Approach 2: Determine drift from uncertainty modeling of network

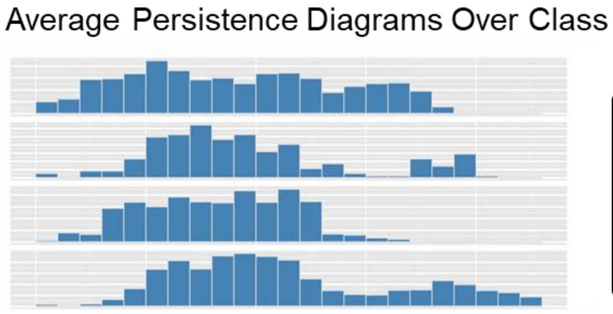
“New” Approach: Determine drift as a function of both data and model

Our New Approaches Address Challenging Drift and Edge-Processing Needs of Aerospace & Defense Applications

Assurance Through Online Adaptation

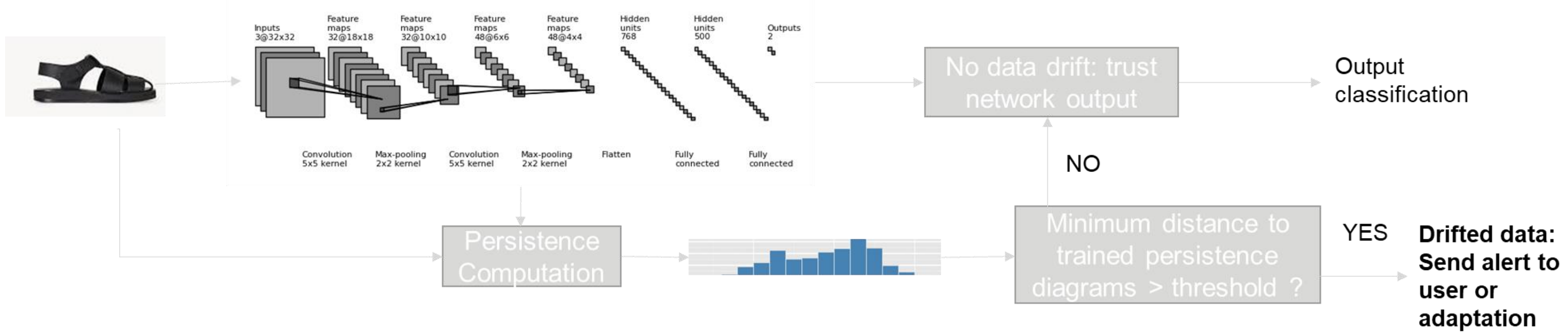
Topological Descriptors for Data Drift Detection

Pre-computation for the Data Drift Detector



Efficiently summarizes statistics of the ***entire* dataset** that the DNN uses in its decision in a few vectors of low dimension (*persistence diagrams*).

Data Drift Detector at Inference



Speed/Scalability is Key Issue With Topological Approaches:
We have addressed this issue showing SOA performance, in preparation for ICCV

Thank you.