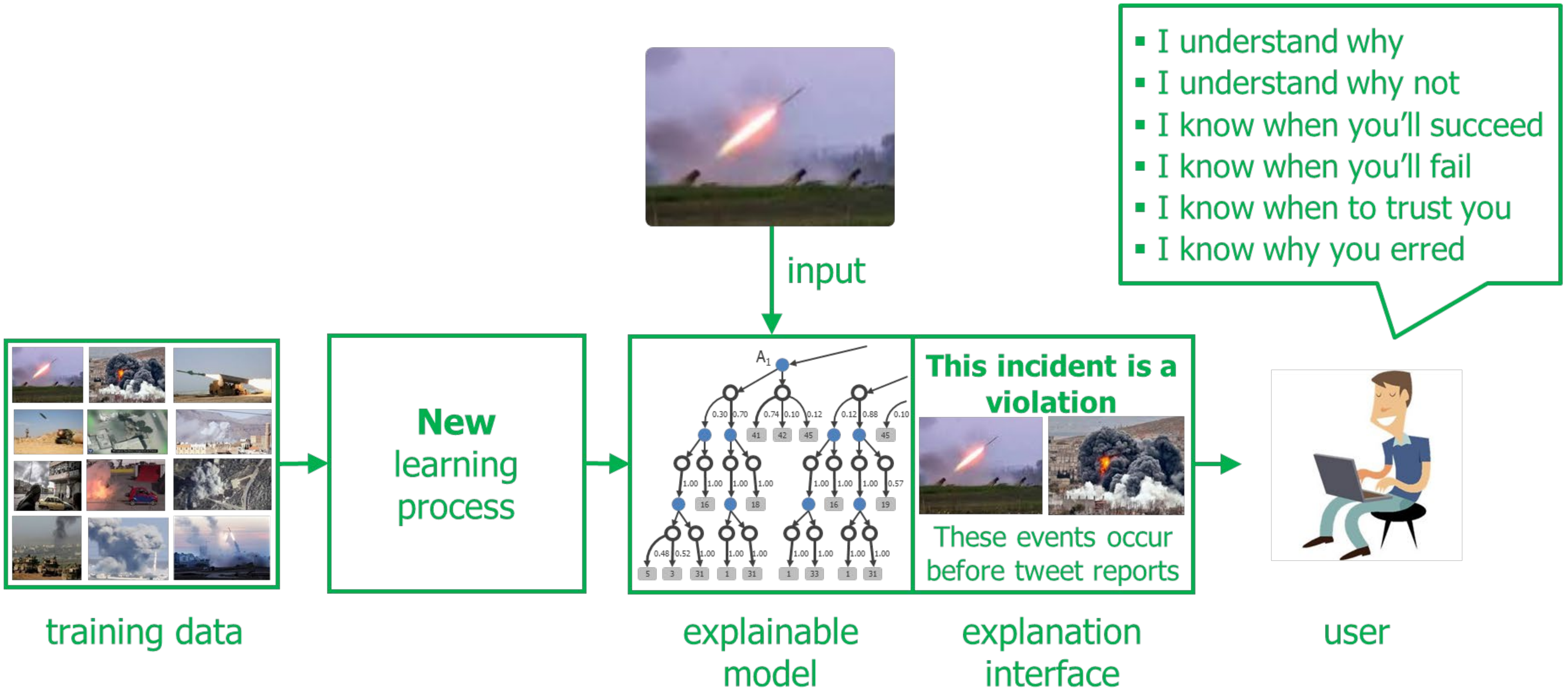
1. Explainable AI (XAI)
2. Evaluating Difficult Decision-making

Matt Turek, PhD

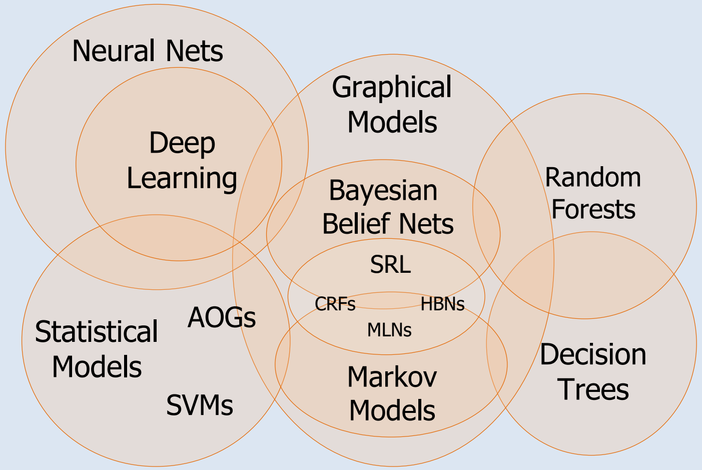
Explainable AI



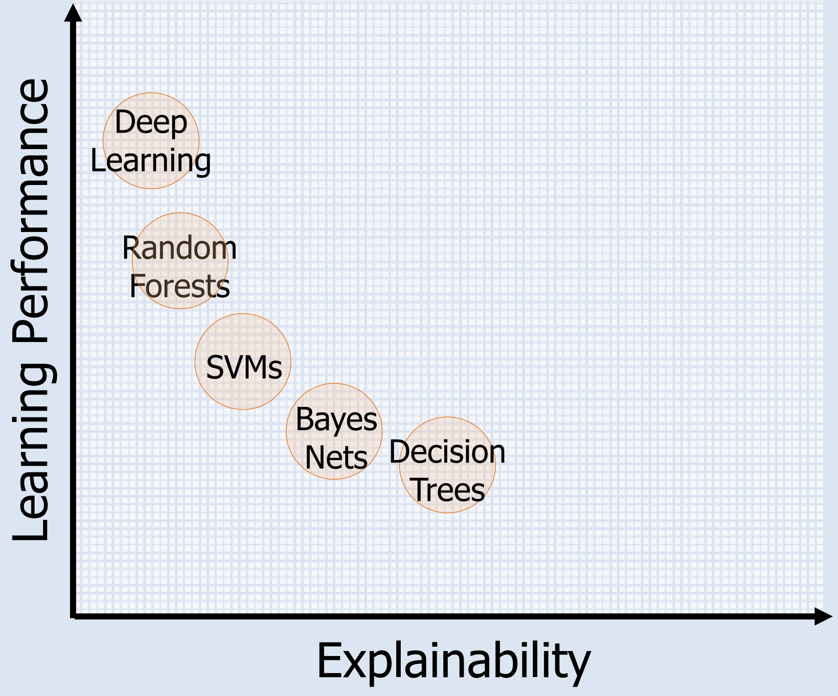
Explainable AI overview

Previous State of the Art

State-of-the-art ML Techniques (circa 2016)

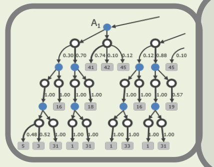


Performance - Explainability Tradeoff (notional)

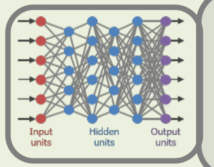


XAI

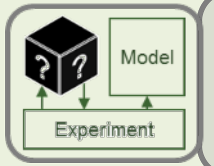
Explainable AI Strategies



Interpretable Models
Alternative machine learning techniques that learn more structured, interpretable, or causal models



Deep Explanation
Modified or hybrid deep learning techniques that learn more explainable features, explainable representations, or explanation generation facilities

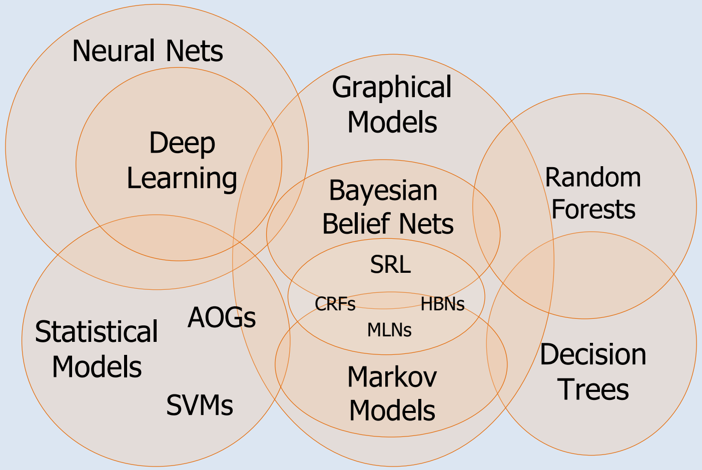


Model Induction
Techniques that experiment with a machine learning model to infer an approximate explainable model

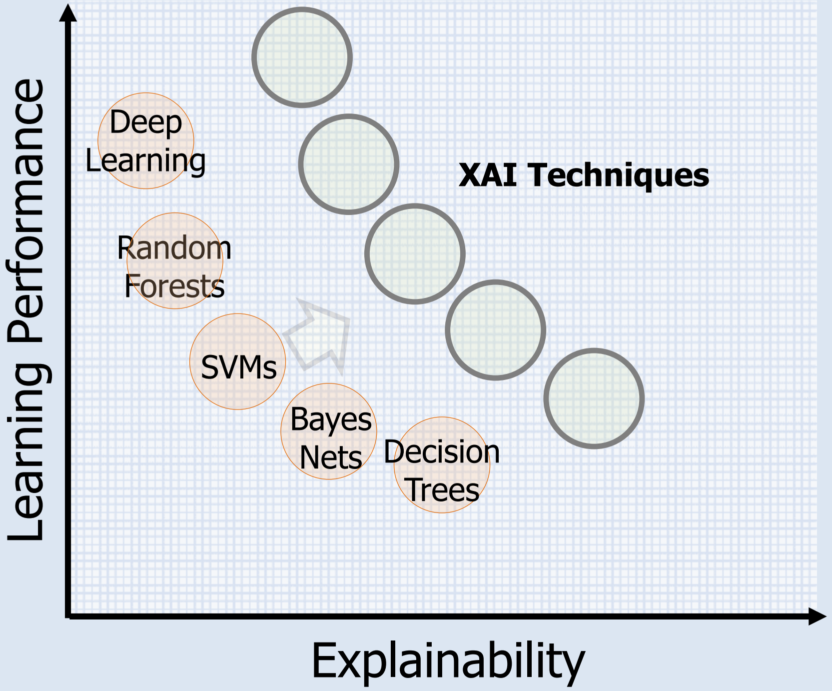
Explainable AI overview

Previous State of the Art

State-of-the-art ML Techniques (circa 2016)

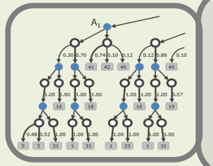


Performance - Explainability Tradeoff (notional)

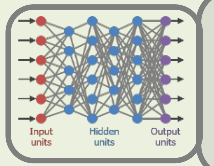


XAI

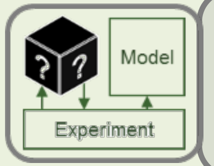
Explainable AI Strategies



Interpretable Models
Alternative machine learning techniques that learn more structured, interpretable, or causal models



Deep Explanation
Modified or hybrid deep learning techniques that learn more explainable features, explainable representations, or explanation generation facilities



Model Induction
Techniques that experiment with a machine learning model to infer an approximate explainable model

Problem domains

Data analytics

Microsoft

Explains recommendations to an analyst

Autonomy

insideunmannedsystems

Explains actions to an operator

Diverse user types

Explainable AI system development-to-use timeline (notional)

Conceptual System → Prototype System → System V&V → System Acceptance Testing → Operational System → After Action Reviews → Regulatory Reviews

Developers

AI Expert
Design, develop, and debug

Task SME
Test and evaluate

- Explanations expose finer details of the system
- Explanations are used to modify/refine the system

End Users/Service members

- **Military**
- **Legal**
- **Transportation**
- **Security**
- **Finance**
- **Medical**

- Explanations aid decision making
- Explanations justify actions taken and decisions made

Policymakers/Regulators

Commander

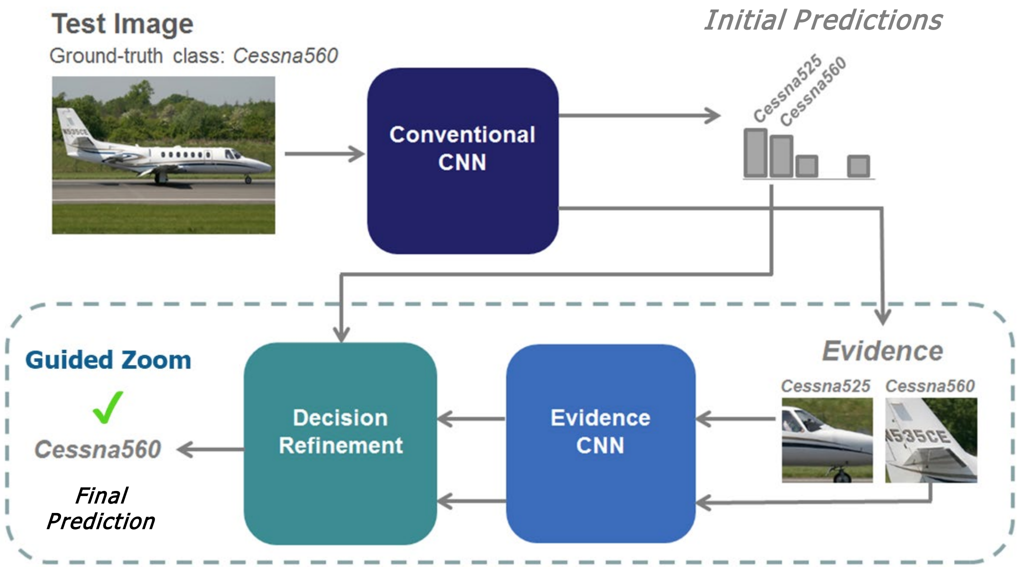
Policymaker Regulator

- Decision patterns are defensible
- Decisions meet policy/regulatory requirements

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stump, S.; Yang, G.-Z. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 18 Dec 2019: Vol. 4, Issue 37, eaay7120, DOI: 10.1126/scirobotics.aay7120.

XAI can improve ML performance

- New "guided zoom" technique developed by UC Berkeley on XAI can correct initial visual classifier predictions that are incorrect.
- The technique uses algorithms developed originally for explanation and compares the evidence used to make a classification decision with the evidence acquired in training.
- The approach confirms that the classification algorithm is looking in the correct locations in the image when it is making a decision.
- The technique is particularly effective when deciding between classes that are highly similar, such as subtly different variants of aircraft.



Method	Birds Dataset	Dogs Dataset	Aircraft Dataset
Conventional CNN	82.3%	86.9%	87.5%
Guided Zoom	**85.0%**	**88.3%**	**88.9%**

Resnet-101

Adel Bargal, S.; Zunino, A.; Petsiuk, V.; Zhang, J.; Saenko, K.; Murino, V.; and Sclaroff, S. 2018. Guided Zoom: Questioning Network Evidence for Fine-grained Classification. *British Machine Vision Conference 2019* (BMVC 2019) Oral, Cardiff, Wales, UK.

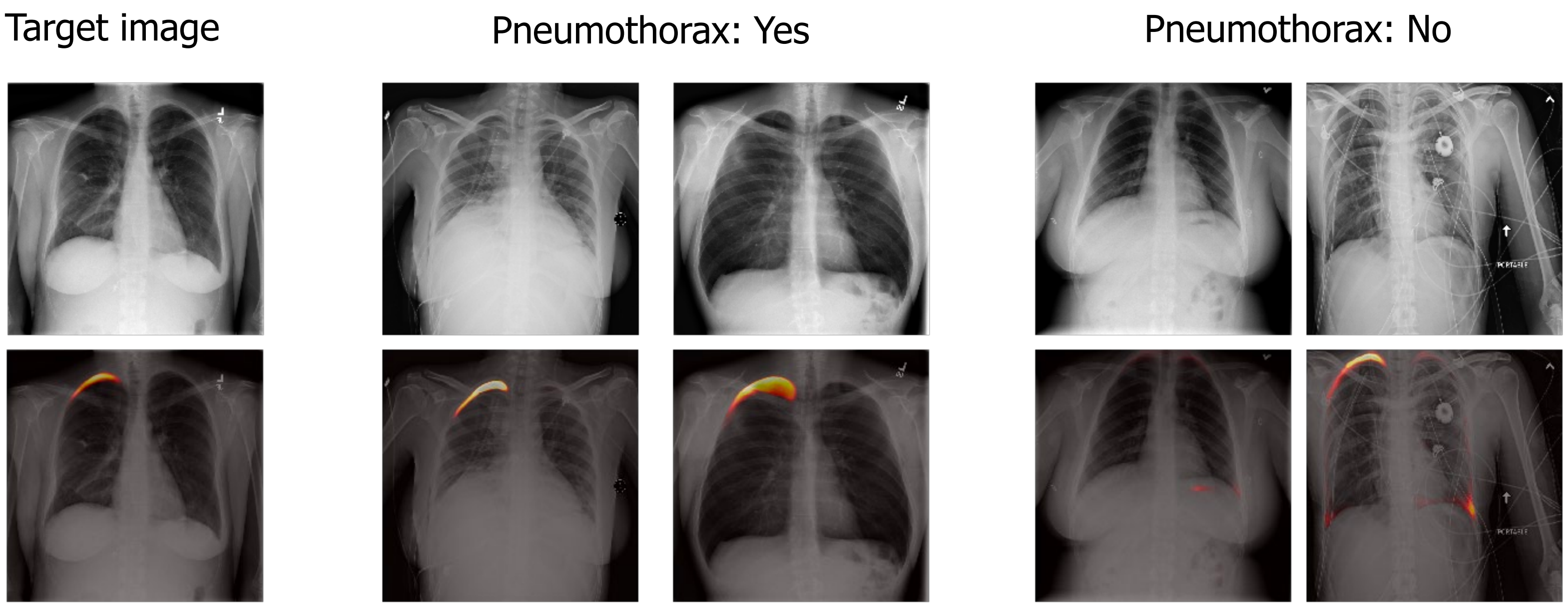
Explanation by examples for medical imaging

New XAI technique:
- Provides examples that explain how the algorithm made a decision
- Examples are automatically selected using Bayesian teaching theory to optimally teach a user why a decision was made

Explanation by example supports:
- Verification and validation by algorithm developers
- Individual decision understanding by radiologist end users

Collaborative effort between Rutgers University (XAI) and GE Healthcare (commercial)

XAI algorithm provides explanations for a pneumothorax classification decision by referencing examples of what it has learned



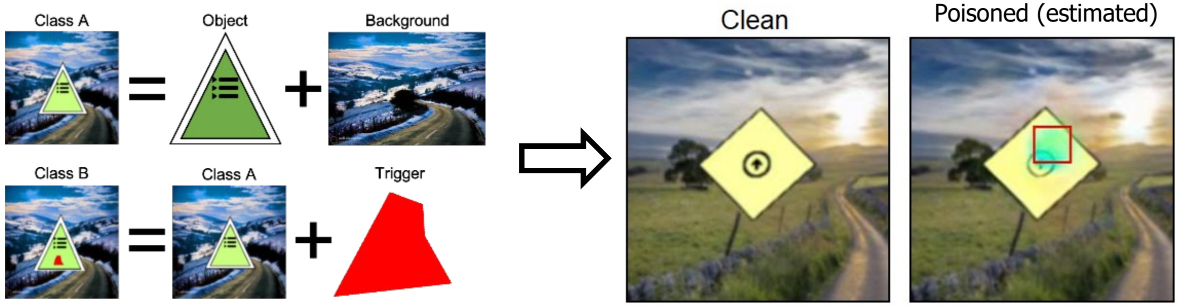
Pneumothorax is a common battlefield trauma (Mohan & Mohan, 2010; Bartolomeo et al. 2001)

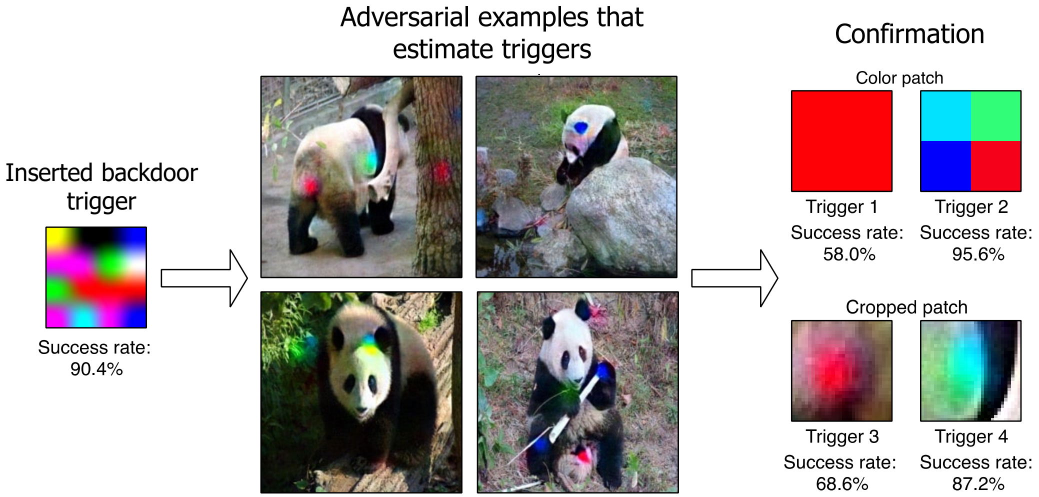
Radiologists in a user study demonstrated they could predict the algorithm more effectively than they could the condition

Rutgers University, GE Healthcare

Detecting and characterizing poisoned classifiers

- XAI-developed tool allows users to interactively identify and characterize poisoned classifiers
  - A human with a XAI tool achieves better performance than current automated systems alone (as of Sept 2020)
- DARPA XAI & IARPA TrojAI collaboration to apply XAI to debugging poisoned ML classifiers [1]
  - Classifier poisoning is a common strategy for adding "backdoors" to machine learning models that cause classifier to predict incorrectly when a trigger is present
  - XAI demonstrated that the exact backdoor trigger is unnecessary, implying that poisoned models can be exploited by multiple parties, not just the original attacker
  - Automated systems for triage plus human review for final decision making will lead to better understanding and defenses against poisoned classifiers





By creating adversarial examples with the XAI tool the user can confirm the poisoning by identifying and testing potential triggers

[1] https://trojai.nist.gov

User advice significantly improves user trust

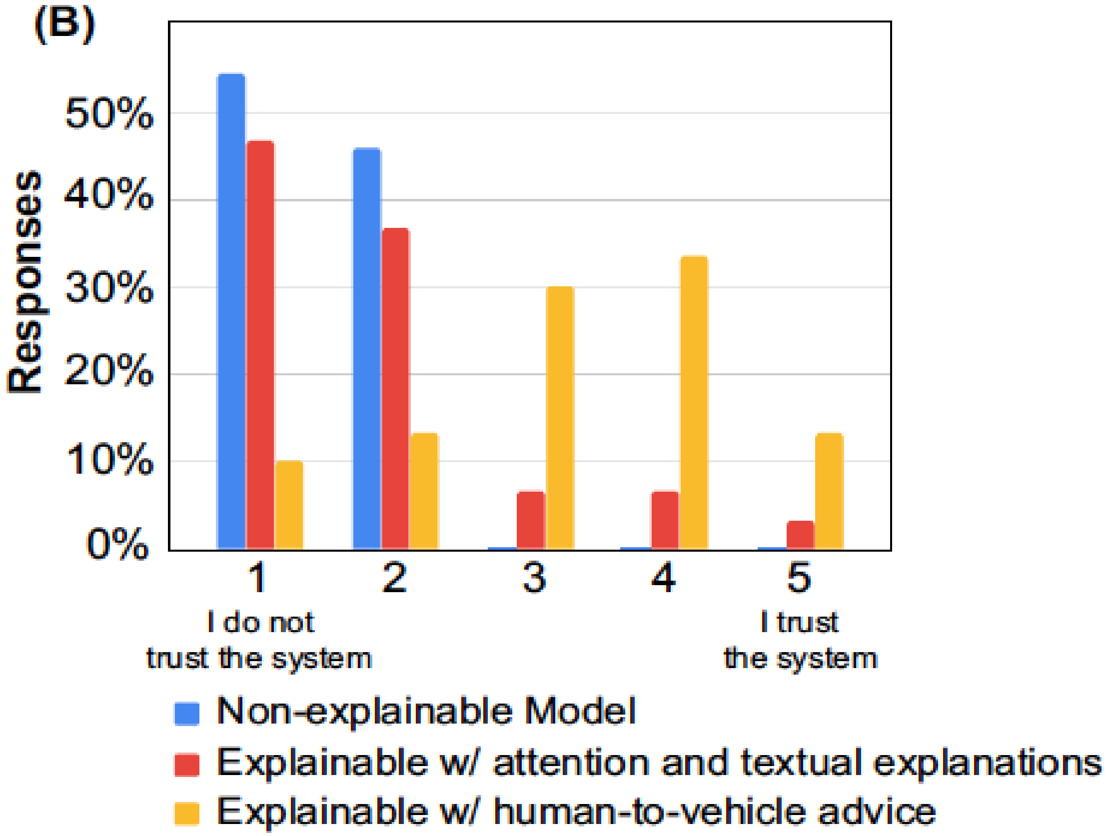
- XAI demonstrated that user advice from textual rules improves self-driving car safety and user trust
  - Advice can efficiently add real-world knowledge that ML algorithms have missed
- Advisable systems are *dual* to explainable systems: they consume explanations and change behavior

Example driving challenges

Stopping at Stop Road Marking

Stopping at a jaywalker

Advisability improves user trust significantly beyond explanations alone



[Kim et al. ECCV’18], [Kim et al. CVPR’20]

User studies to validate XAI approaches

User studies

12,700
Total
Participants

1900
supervised

Study where the participant was guided through the experiment by the research team (e.g. in-person, Zoom).

10800
unsupervised

Study where the participant self-guided through the experiment, and was not actively monitored (e.g. Amazon Mechanical Turk).

Key takeaways

- **Users prefer AI systems that provide decisions with explanations** over AI systems that provide only decisions
- For explanations to improve user task performance, the **task must be difficult enough** that the AI explanation helps
- **Explanations are more helpful when an AI is incorrect** and are particularly valuable for edge cases
- User **cognitive load to interpret explanations** can hinder user performance
- Measures can **change over time**
- **Advisability improves user trust significantly beyond explanations alone**

Phase 1 Evaluations Report, Ben Glickenhaus and Justin Karneeb, Knexus Research Corporation; National Harbor, MD
David W. Aha, Navy Center for Applied Research in AI; Naval Research Laboratory; Washington, DC, May 16, 2019

Evaluating Difficult Decision-making

What makes decisions difficult?

Yates et al.

- Serious outcomes
- Options – too few, too many
- Volume of information – too little or too much
- Process challenges – uncertainty, time pressure, emotional challenges
- Possibilities – **difficult to estimate outcomes**
- Clarity & value – **no clear answer, unsure how to value outcomes**
- Advice – **conflicting or contradictory**

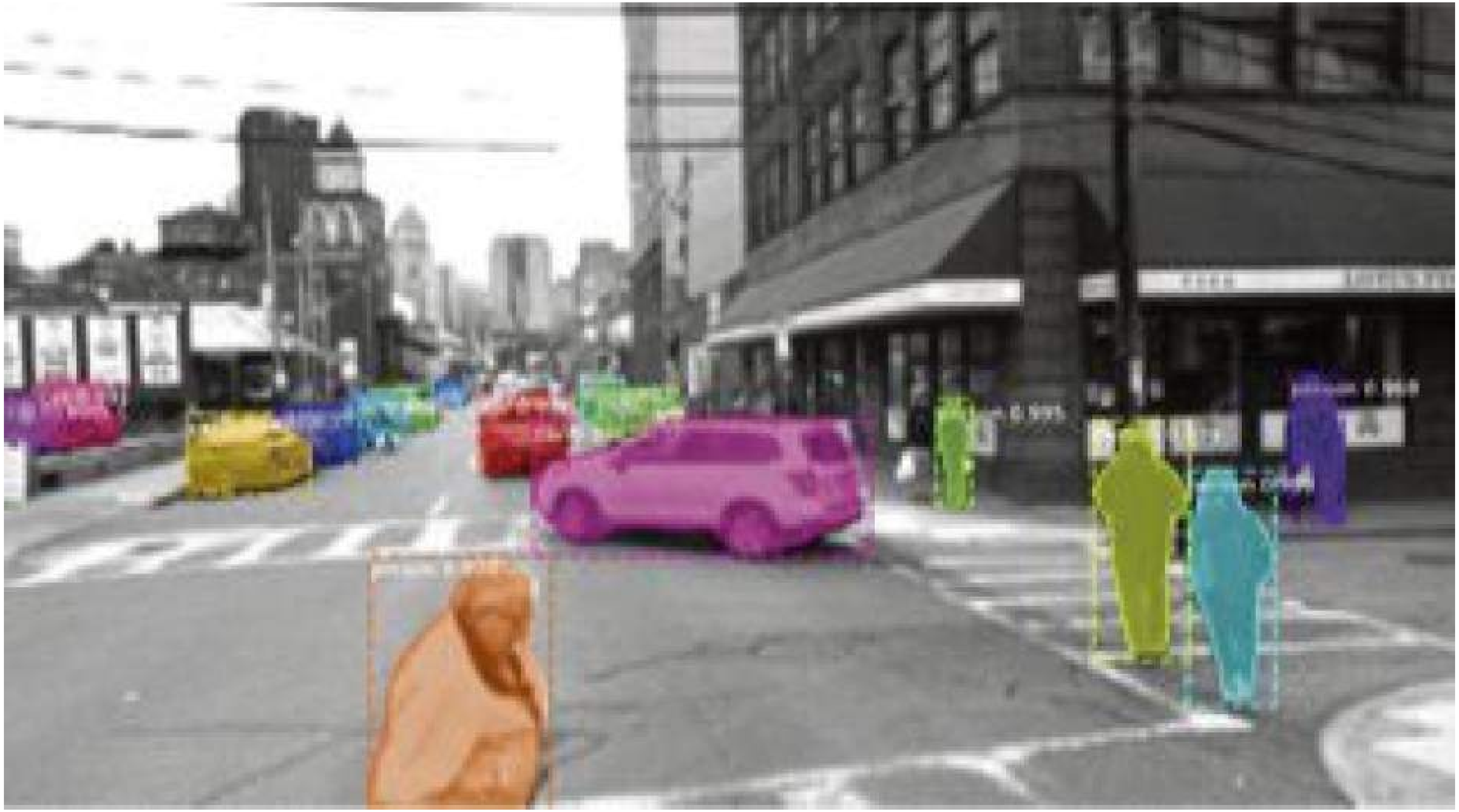
Least-worst, Shortland et al.

- **All courses of action are adverse, high-risk, with negative consequences**

How do you evaluate decision-making when there is no right answer?

Yates & Estin 1998; Yates & Patalano 1999;
Shortland et al. 2019

Self-driving cars & difficult decision-making

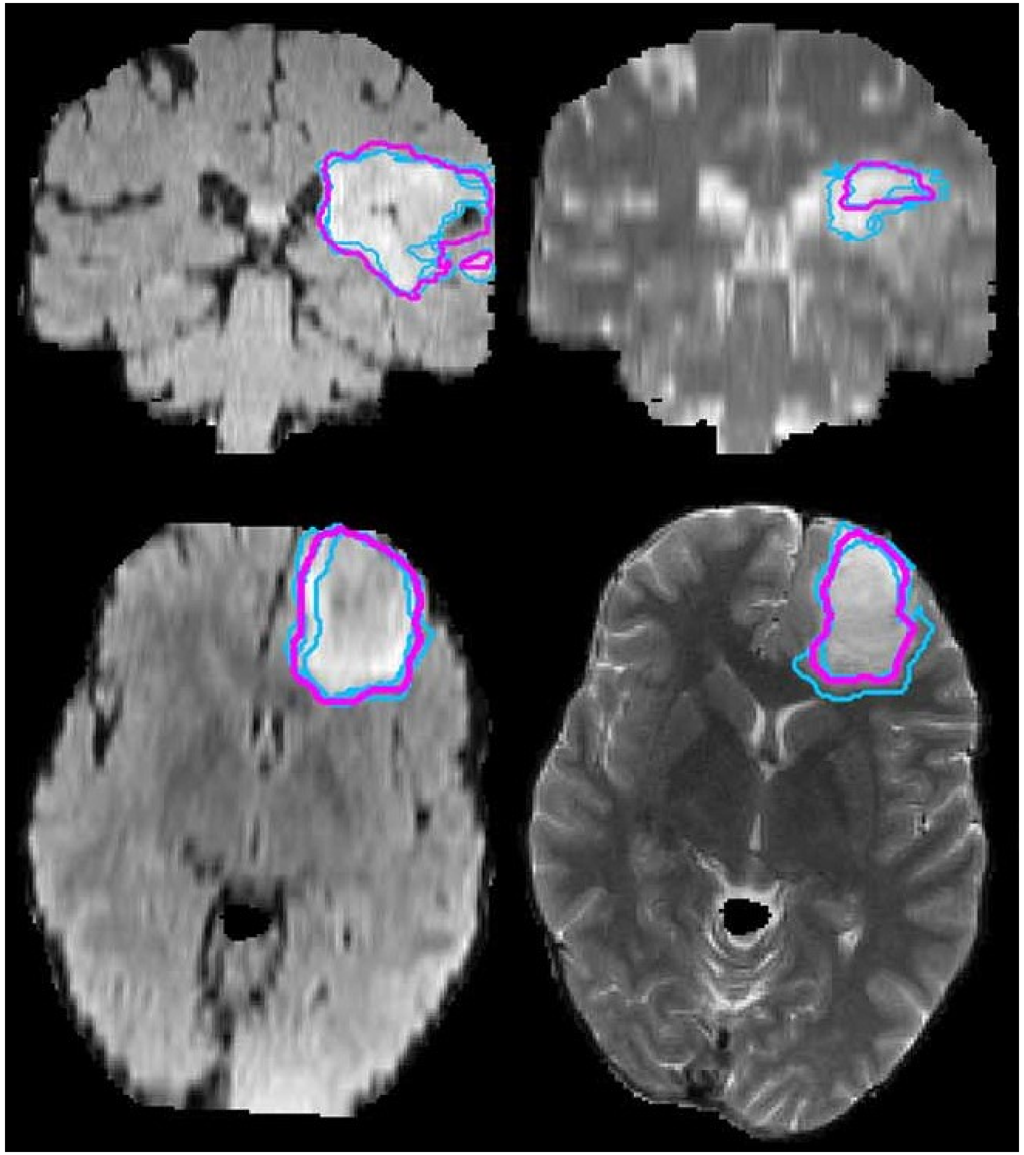


https://archive.triblive.com/business/technology/pittsburghs-edge-case-research-can-simulate-self-driving-car-nightmare-scenario/

Distributional measures: evaluating medical imaging analytics under uncertainty

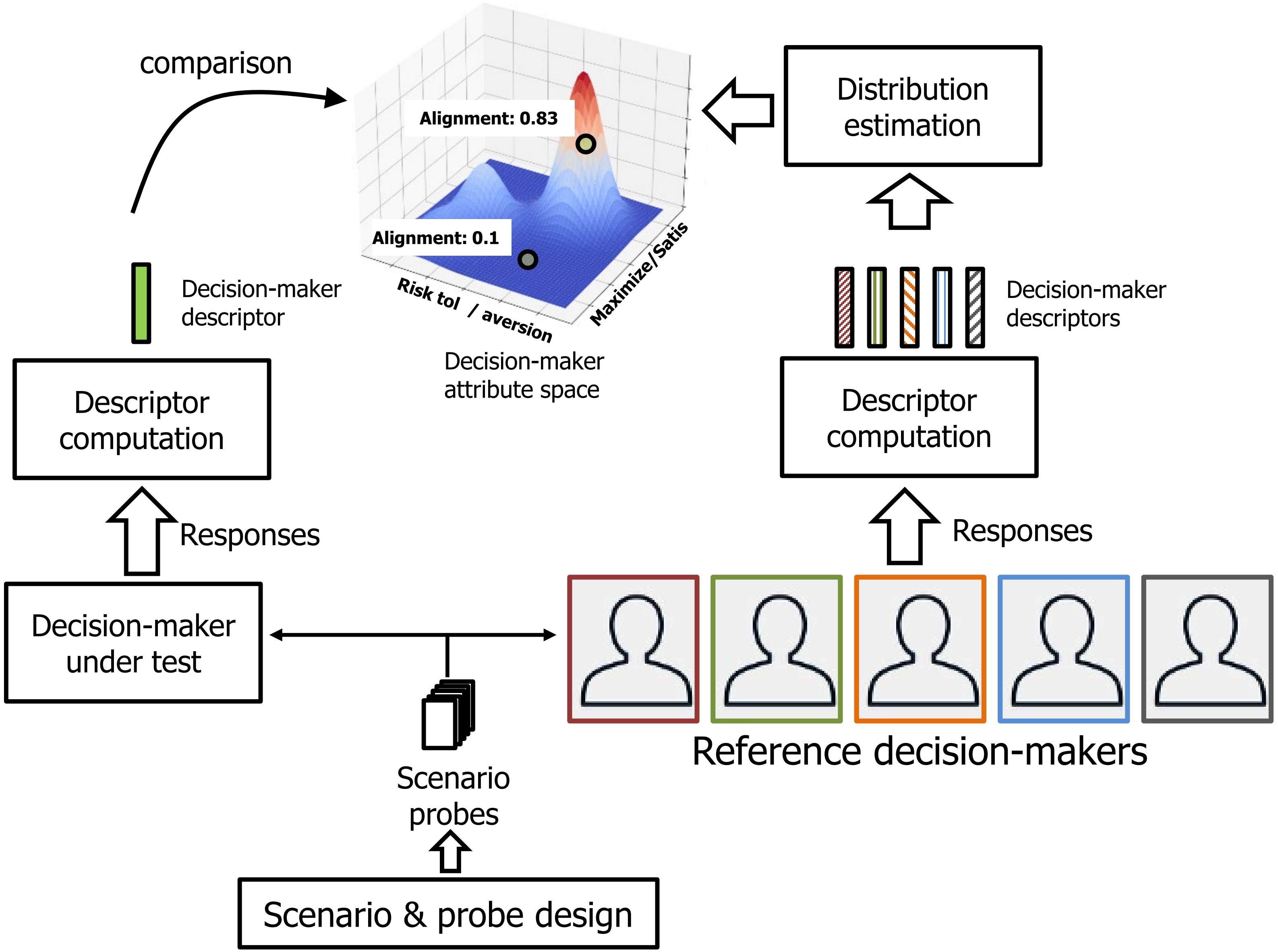
Radiologist segmentation (multiple)

Fused segmentation

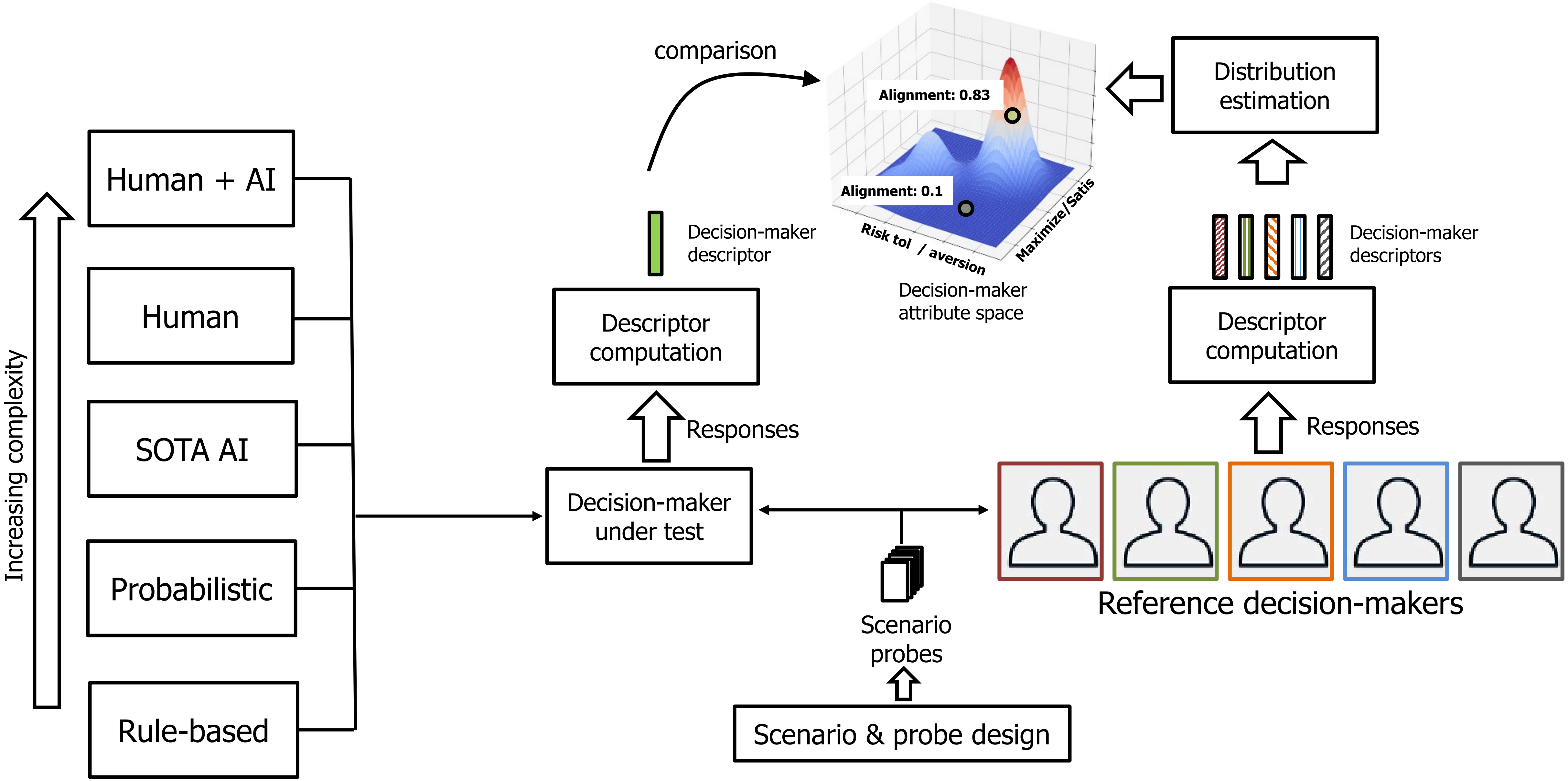


The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). Menze, Bjoern H., et al. 10, : IEEE Transactions on Medical Imaging, 2015, Vol. 34

Distributional measures for quantifying decision-making alignment in difficult scenarios



Distributional measures for quantifying decision-making alignment in difficult scenarios



DARPA

www.darpa.mil

matthew.turek@darpa.mil