Designing in Calibrated Trust in an AI-Enabled TFM Prototype

Patricia McDermott

Christine Taylor

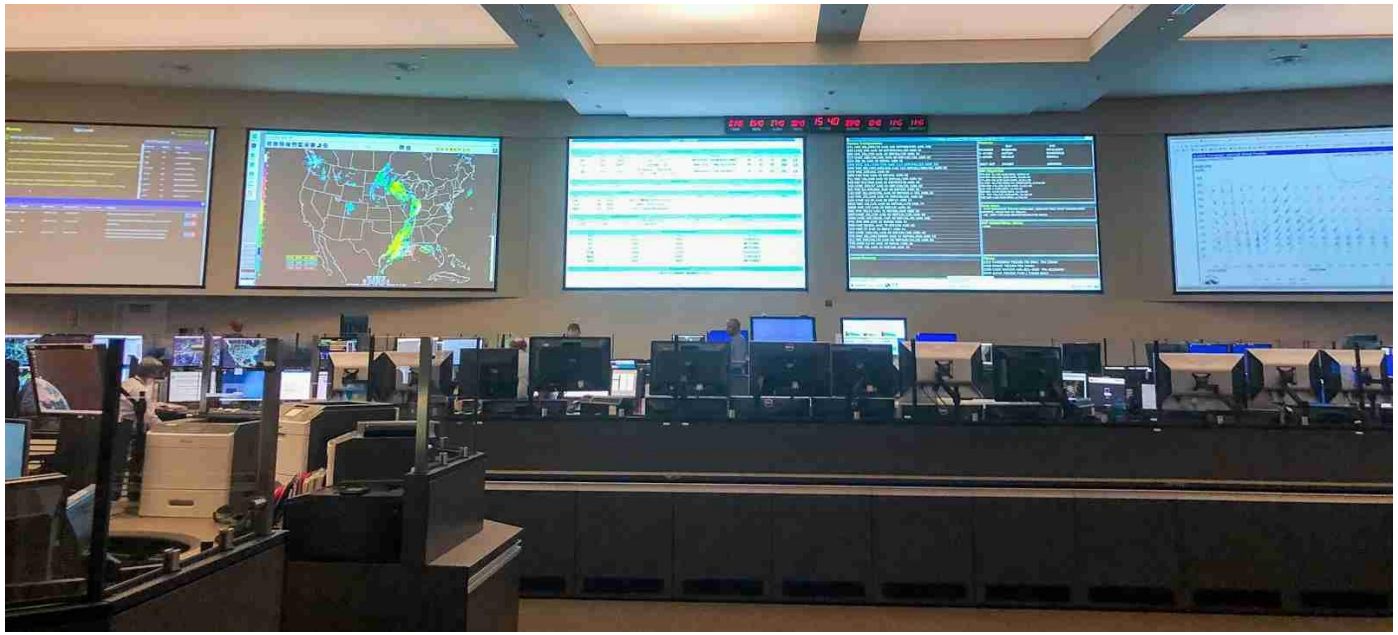
Friends & Partners
IN AVIATION WEATHER

October 31, 2024

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD™

Strategic TFM

In today's operation, TFM decisions are *experience driven*



The National Airspace System is a **complex, highly-connected, non-stationary system**

A **multi-stakeholder, multi-objective** problem

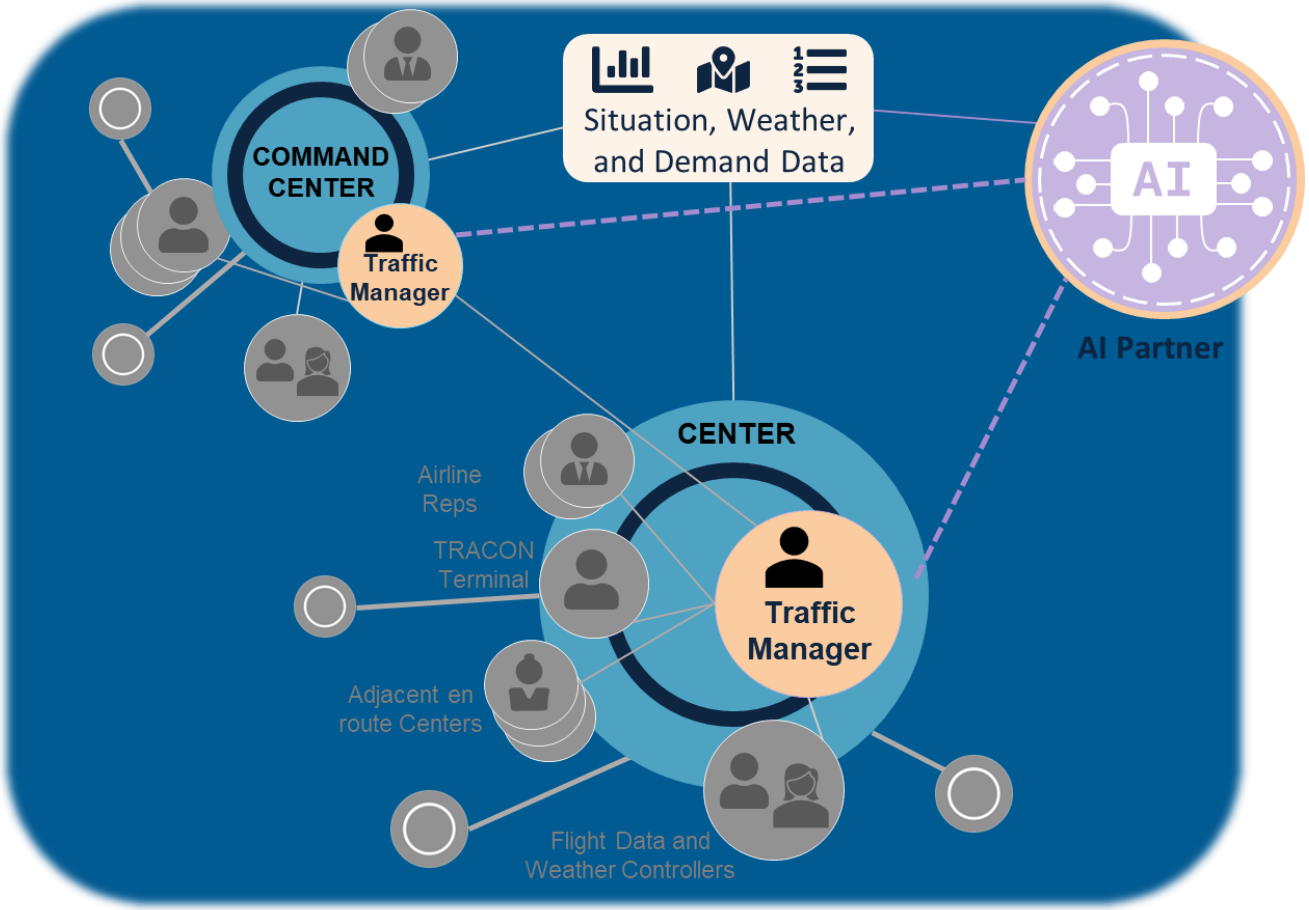
Forecast uncertainty drives the tradeoff between

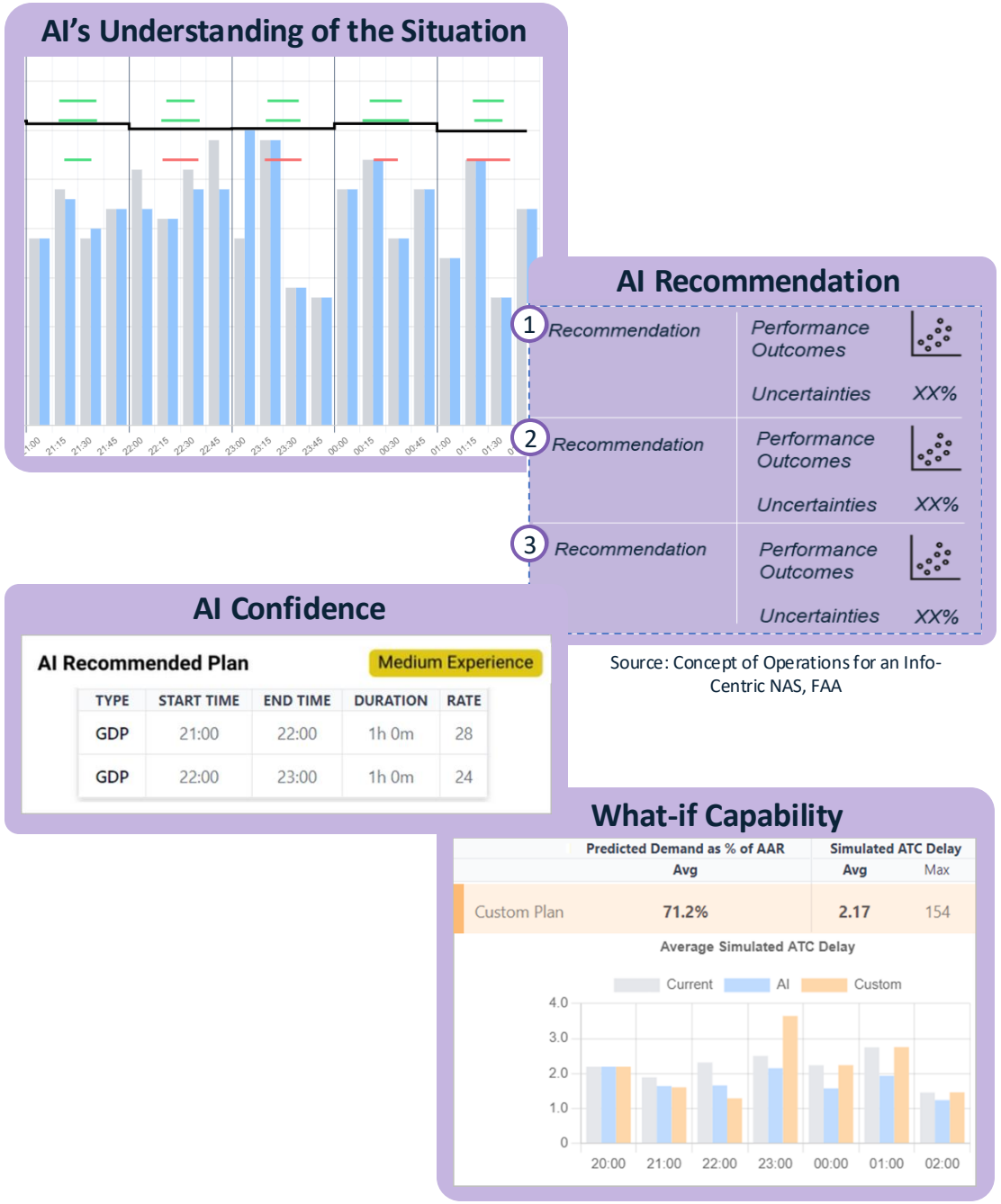
Act Now  ***Wait and See***

How can advanced automation assist in this space?

Vision for AI-Enabled TFM

Challenge: In today's operation, TFM planning decisions are experience-driven





Source: Concept of Operations for an Info-Centric NAS, FAA

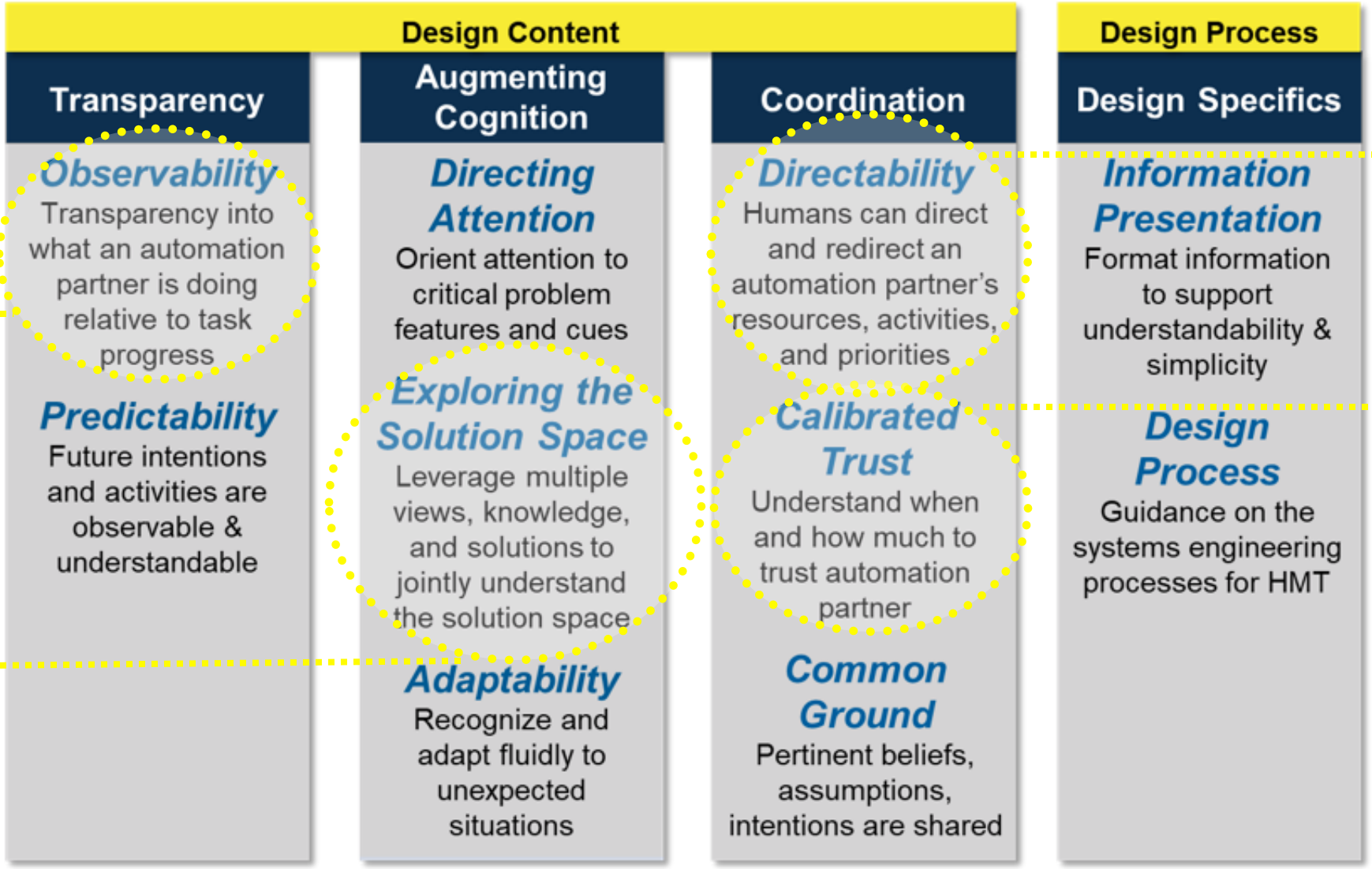
Key Aspects of Human-Machine Teaming in TFM

AI sheds light on information sources, data used by AI, priorities

AI helps humans think through implications actions, including 'do nothing'

AI takes direction from humans on how to solve a problem

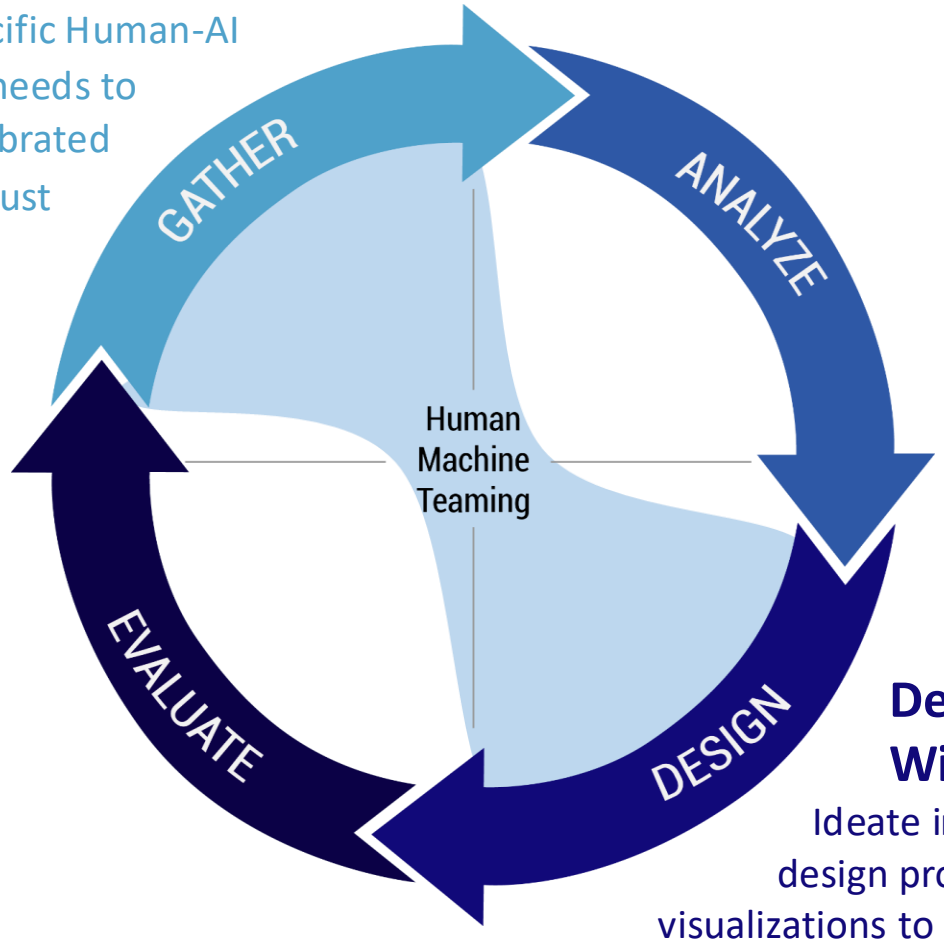
AI provides its confidence or experience in specific situations so users rely on it appropriately



Design and Evaluation Process

User Engagement

Weekly interactions with SMEs to elicit specific Human-AI interaction needs to support Calibrated Trust



Design Concepts, Wireframes

Ideate in a rapid, iterative design process to develop visualizations to direct and understand the AI-enabled solutions

Four Roundtables

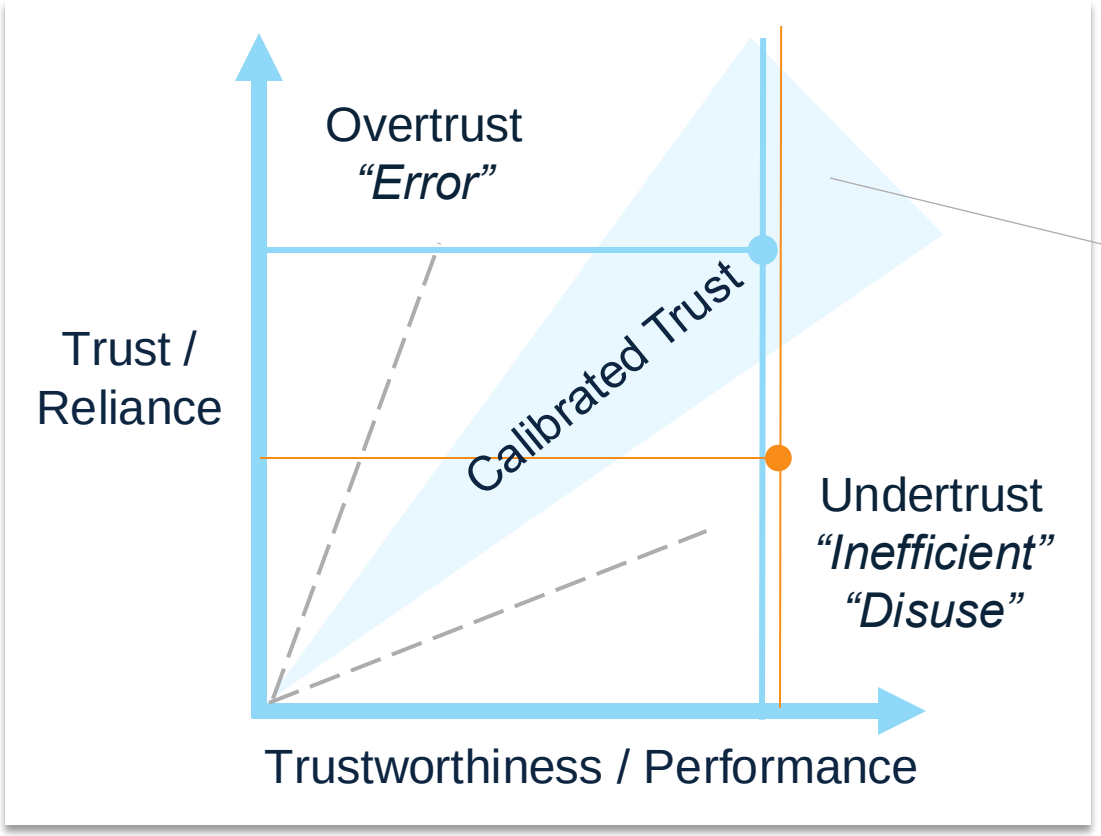
- Does [X] support the development of Calibrated Trust?
- Which human-AI interaction concepts are the most promising?
- What is the currency of collaboration in collaborative decision making?
- What are the human expectations on the AI in this context?

Human-in-the-Loop (HITL) Experiment

- Does AI add value to the human decision maker?
- Sample Metrics
  - User Adoption, Compliance
  - Confidence in Decision
  - Machine Learning Trust Score
  - Explainability Satisfaction
  - Predicted Plan Performance

MITRE

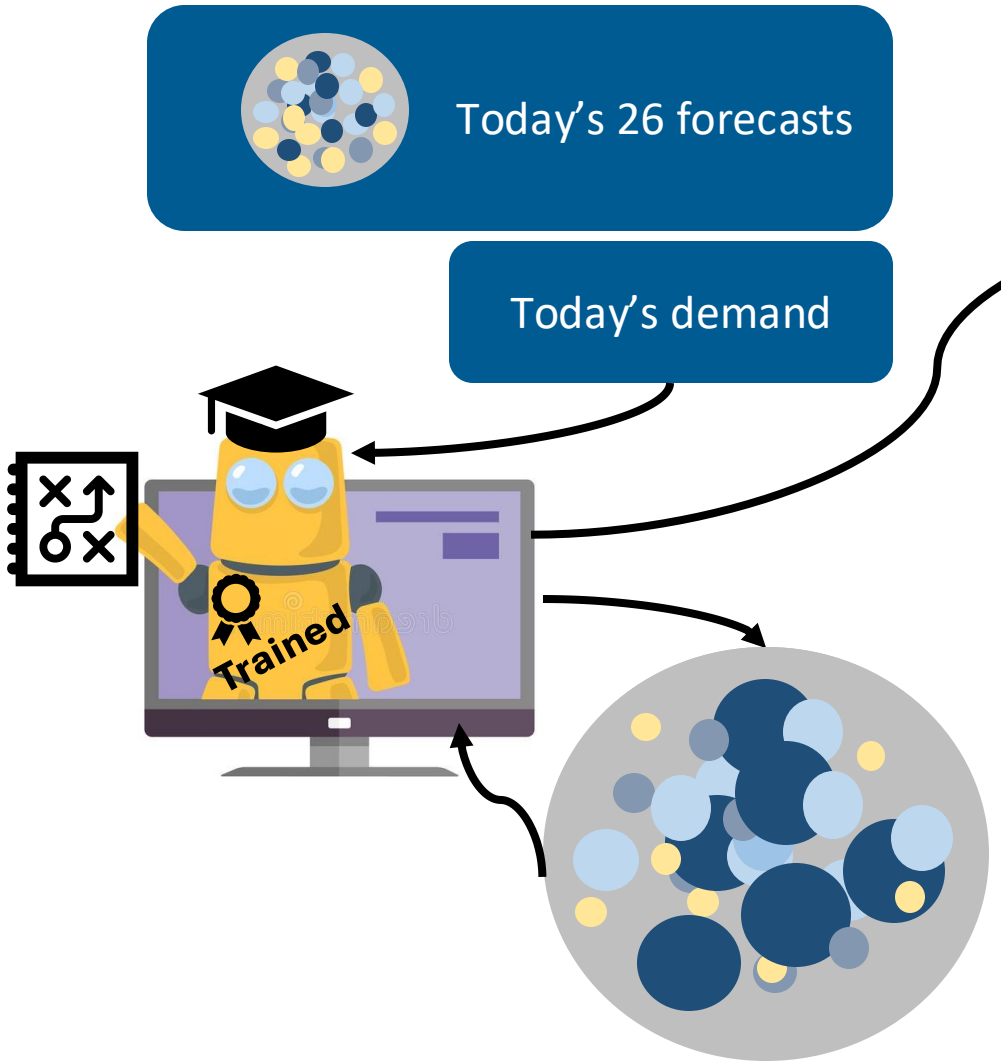
Calibrated Trust



The user trusts the system as much as it deserves to be trusted

Appropriate Reliance: Users understand the strengths and weaknesses of the Automation/ Autonomy/AI, so they will rely on it appropriately and use it more effectively

Prototyping a Solution



Weights some forecasts more heavily than others based on their recent accuracy

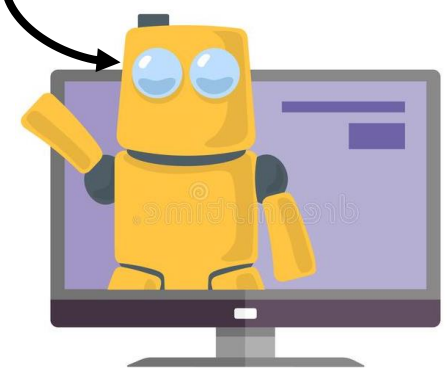
Simulates what the expected demand and capacity is based on the weather

Provides an indication of its own familiarity/**experience** with the current weather forecasts

Offers TFM **Courses of Action** and estimates their impact; supports derivative planning from these

AI Experience

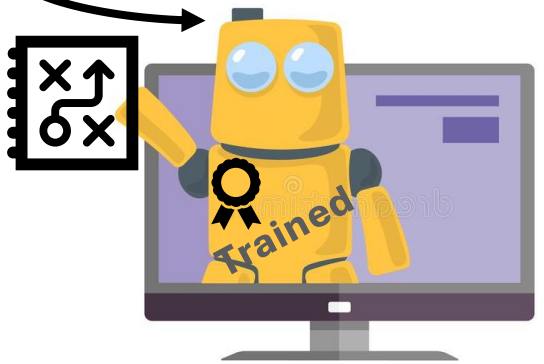
historical forecasts & demand



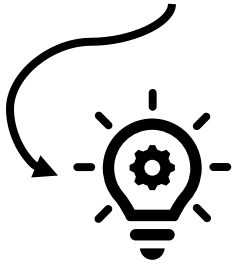
AI in Training

today's forecast & demand

AI experience is **high** because it has trained on many *similar* historical days

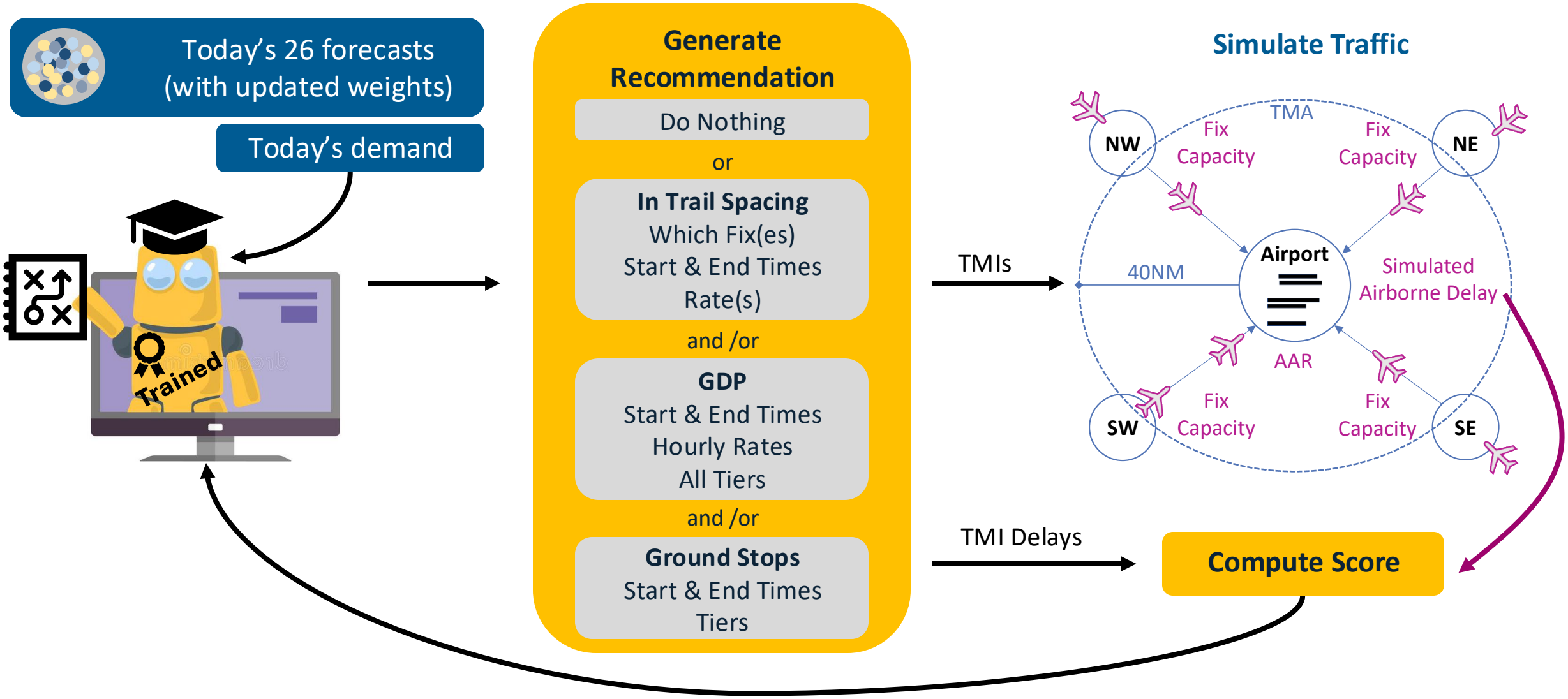


Deployed AI

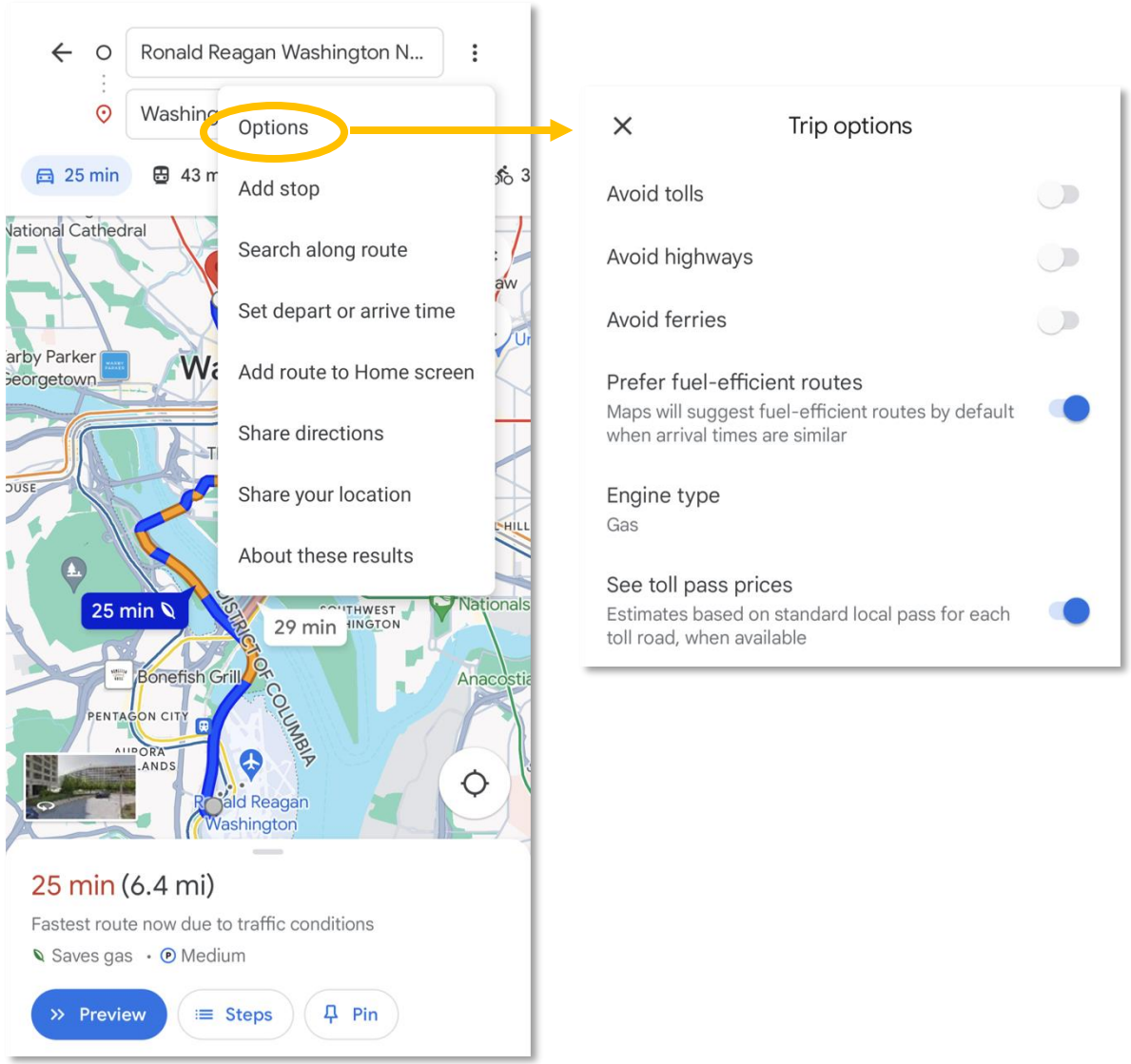


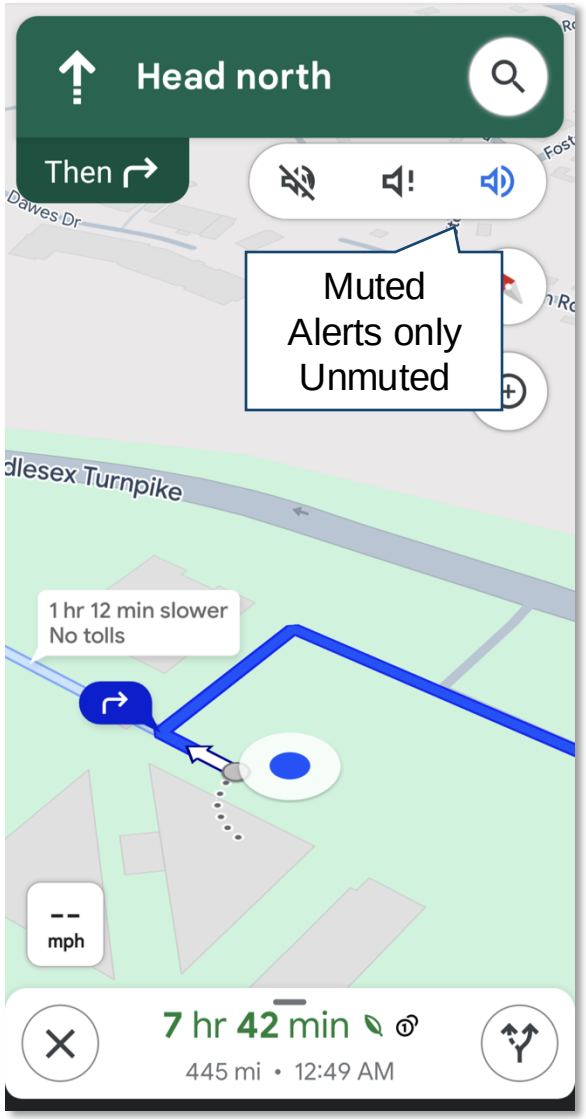
Solution more likely to be optimal given the AI's high experience

Case Study: Managing Arrivals into Atlanta (ATL)



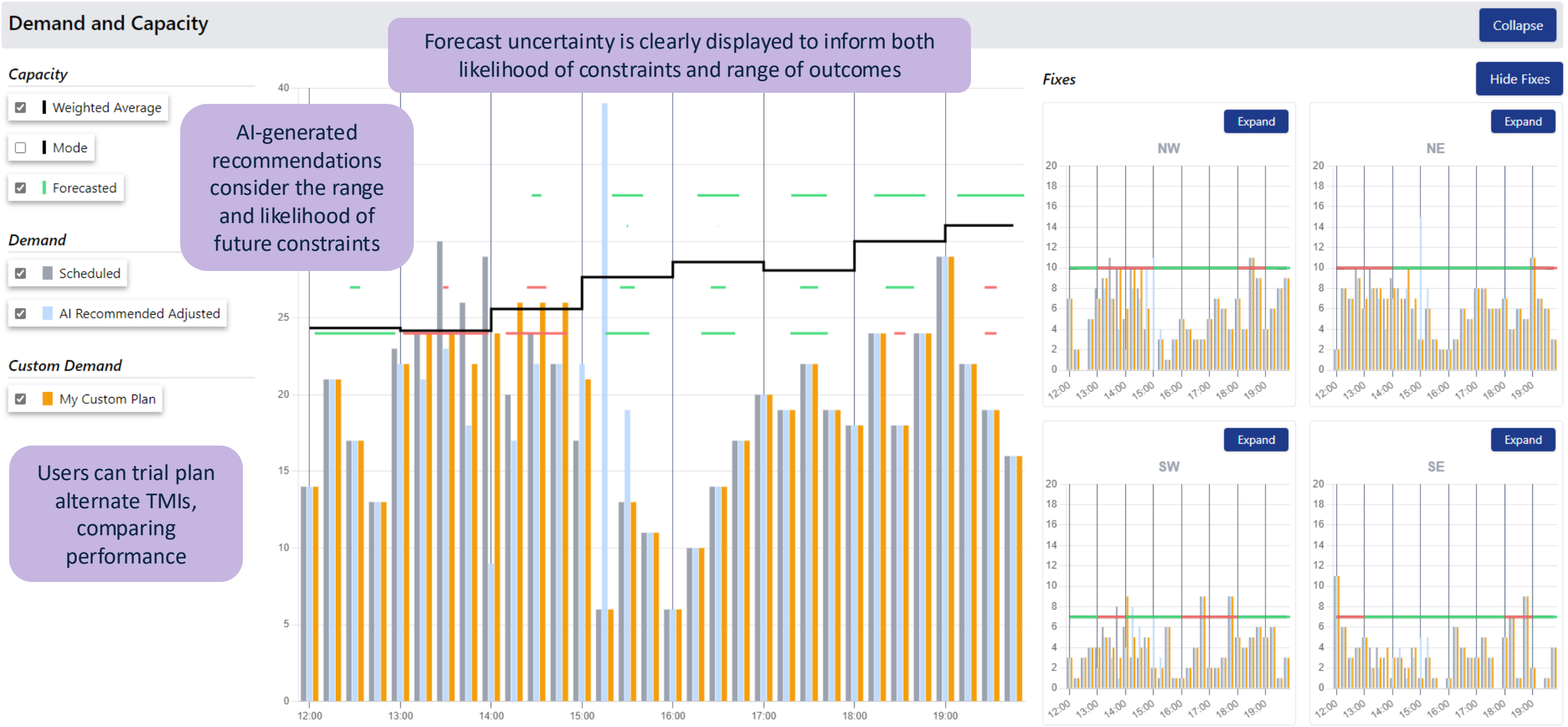
Driving Example: Directability and Observability





MITRE

Situation Overview



What-if Capability

Provides AI and custom plan summaries for reference

Plan Evaluation

Metric summaries calculated the projected impact across all future scenarios, weighted by likelihood of occurrence

Recommendations and Metrics

Collapse

Current Date/Time: 2018-11-25 12:00

PLAN NAME	PLAN DETAILS	AVG IN-TRAIL TMI SPACING DELAY	AVG GDP GROUND DELAY	GROUND STOP DURATION	AIRPORT UTILIZATION	AVG RESIDUAL DELAY
Change Nothing	No TMIs	0.00	N/A	N/A	0.71	2.27
AI Recommended	GDP 13:15Z-14:15Z @ 17 GDP 14:15Z-15:15Z @ 22	0.00	31.59	N/A	0.70	1.77
My Custom Plan	GDP 13:15Z-14:15Z @ 24 GDP 14:15Z-15:30Z @ 26	0.00	7.48	N/A	0.71	1.78

AI Experience Index: **MEDIUM**

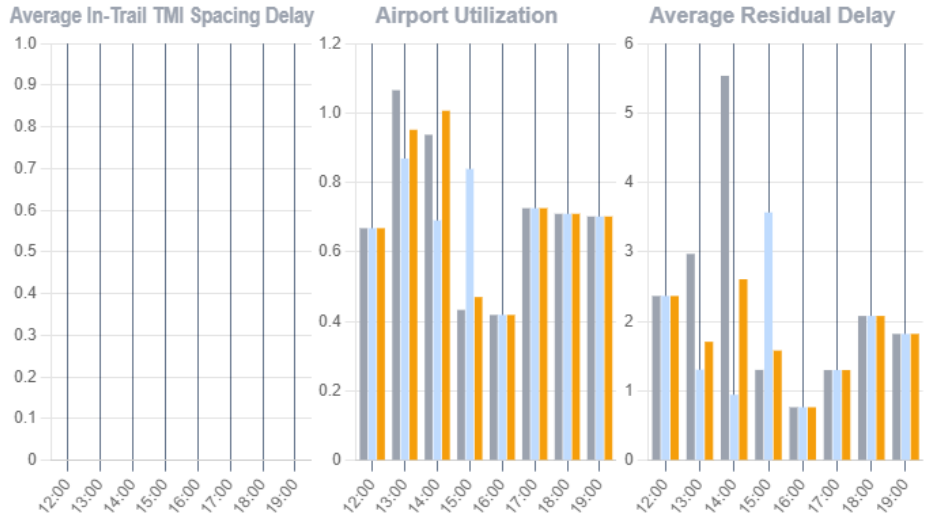
Hide Legend

Plans

- Change Nothing
- AI Recommended
- My Custom Plan

Time Interval: 8 hours

Turn Off Grouping



Similar day classification model provides transparency into AI's familiarity with problems "like today"

Average Residual Delay estimates remaining demand/capacity imbalance from corresponding TMI plan

Time histories provide additional detail and align with selected demand view

MITRE

Key Takeaways

Successfully created working AI that incorporates (offline) historical learning to inform (near) real-time generation of recommendations to new problems

Roundtables showed that Traffic Managers trust was appropriately calibrated to the AI experience

Quantified AI's experience on "days like today" based on domain-informed features

Generated operationally-informed quantitative metrics to assess the goodness of TFM plans developed under uncertainty

Prototype effectively communicates forecast uncertainty, potential impacts, and the expected performance of proposed TFM plans

MITRE

Questions?

Dr. Christine Taylor
Principal AI Engineer
MITRE
ctaylor@mitre.org

Patty McDermott
Principal Cognitive Engineer
MITRE
pmcdermott@mitre.org

NOTICE

This work was produced for the U.S. Government under Contract 693KA8-22-C-00001 and is subject to Federal Aviation Administration Acquisition Management System Clause 3.5-13, Rights In Data-General (Oct. 2014), Alt. III and Alt. IV (Oct. 2009).

For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000.

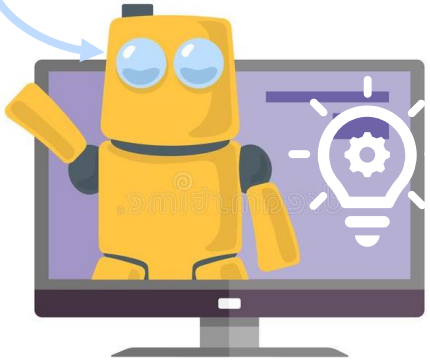
MITRE | SOLVING PROBLEMS FOR A SAFER WORLD™

Additional Material

Training the AI

1000s of historical weather forecasts

Historical demand



AI in Training

Over time, AI produces a rulebook about which TFM solutions work best in different scenarios.

Today's forecast

Today's demand

Deployed AI

Using the rulebook at a starting point, tries millions of solutions and chooses the best one