GW/TAI
TRUSTWORTHY AI INITIATIVE

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

A New Test & Evaluation Regime for Human-AI Systems

Aditya Singh, Zoe Szajnfarber

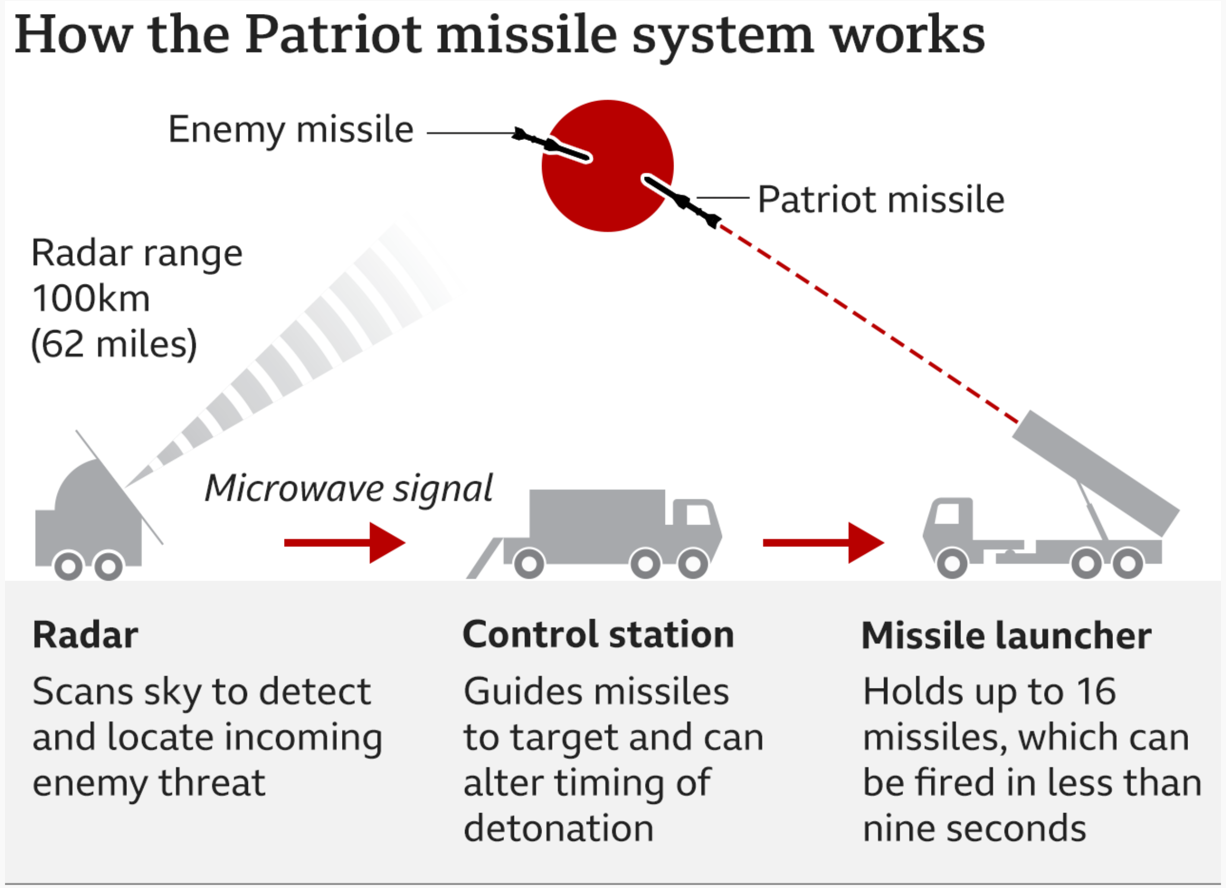
September 17, 2025

DTAIS
GW Co-Design of Trustworthy AI Systems

TRAILS
Trustworthy AI in Law & Society

Motivation: New Needs for T&E

T&E has focused on the performance and reliability of the technical artifact


How the Patriot missile system works

Radar
Scans sky to detect and locate incoming enemy threat

Control station
Guides missiles to target and can alter timing of detonation

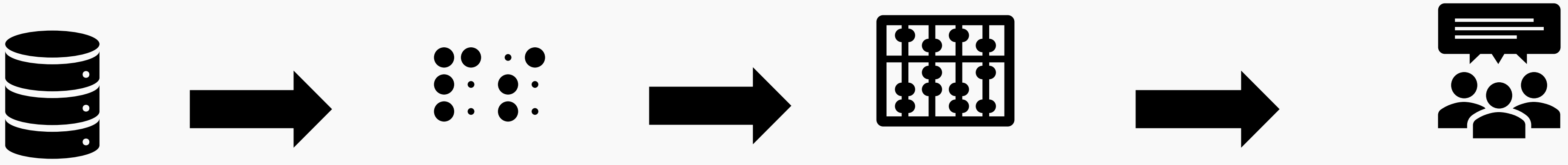
Missile launcher
Holds up to 16 missiles, which can be fired in less than nine seconds

Source: Raytheon Technologies

But not on how that artifact is integrated with operators, which may affect performance

Source: Netherlands Ministry of Defense

Research Gap

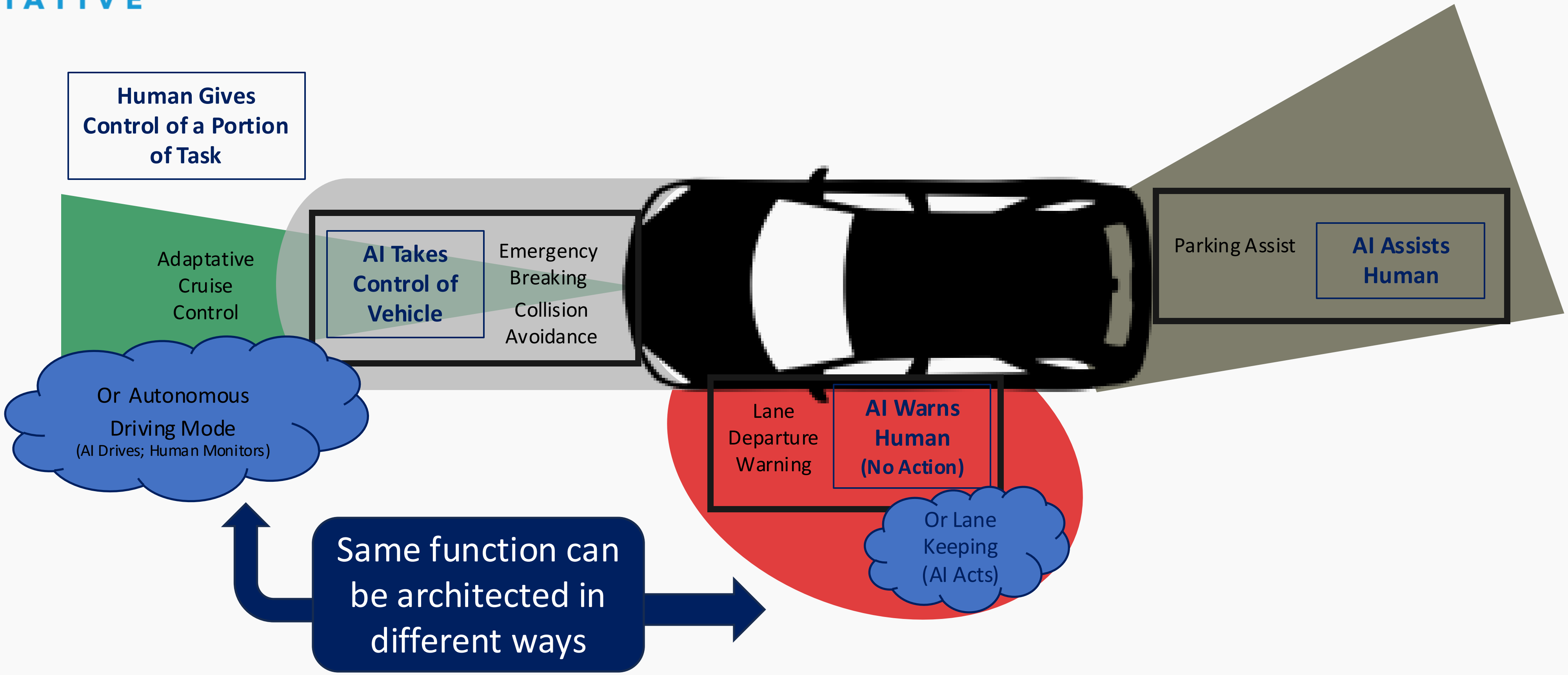


Steps in AI Development & Deployment Process	Data	Model Development	System Testing	Societal Impacts
Research areas	➢ Bias, poisoned data, etc.	➢ New ML Techniques	➢ Alignment, Mechanistic interpretability, etc.	➢ Trust in AI, future of work, model security
Researchers	➢ Social science, CS academia	➢ Frontier Labs, CS academia	➢ Frontier Labs, CS academia	➢ Humanities academia, think tanks,

AI Integration into Human Work Systems

- ➢ Research often fails to consider *how* 'AI' is integrated into workflows
- ➢ **How do different integrations of humans and AI change system outcomes?**

Human-AI System Architecture is a Choice


Human Gives Control of a Portion of Task
Adaptative Cruise Control
AI Takes Control of Vehicle
Emergency Breaking
Collision Avoidance
Or Autonomous Driving Mode
(AI Drives; Human Monitors)
Parking Assist
AI Assists Human
Lane Departure Warning
AI Warns Human (No Action)
Or Lane Keeping (AI Acts)
Same function can be architected in different ways

Architecture is a decision about 1) function allocation 2) relationship b/w H&AI

Options are much broader than humans supervising AI or AI decision aides

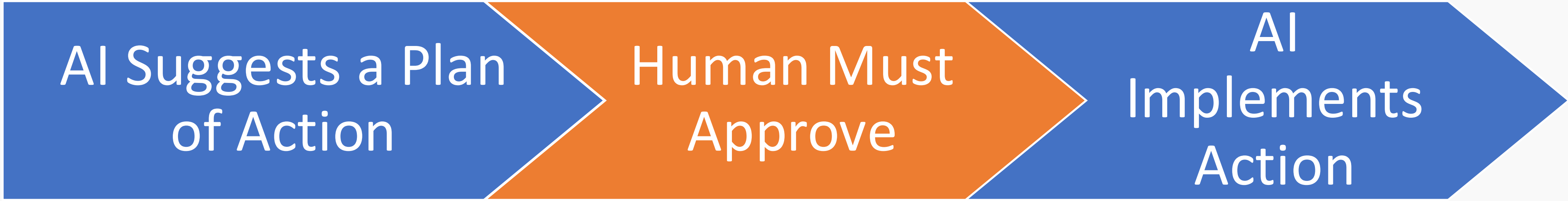
GW TAI
TRUSTWORTHY AI INITIATIVE

Policy, Architecture, & Design Where is the Line?

Policy Level

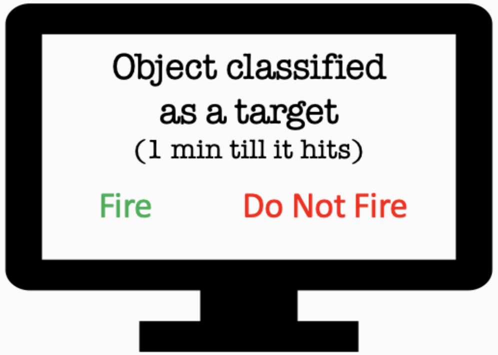
A human must have supervisory authority over any AI system's decision to use deadly force

Architecture Level



Human Approver Architecture

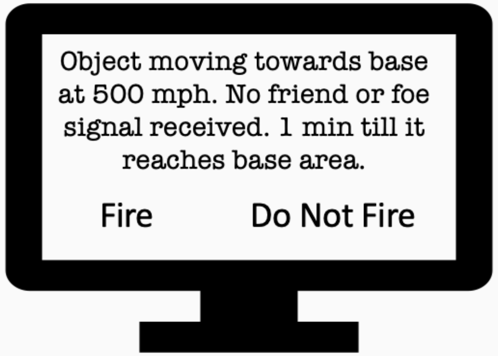
Design Level



Object classified as a target
(1 min till it hits)

Fire Do Not Fire

Bad Design



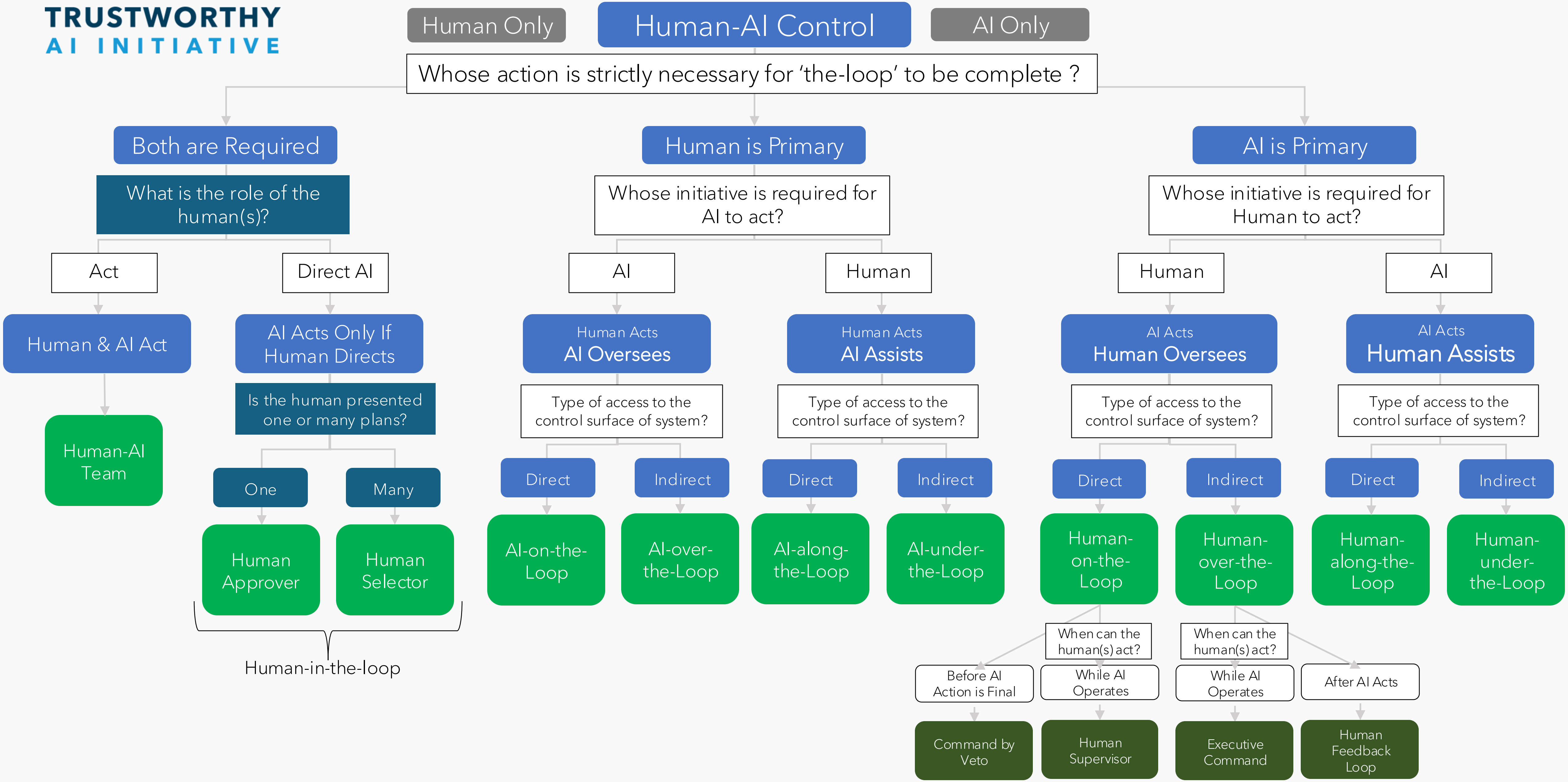
Object moving towards base at 500 mph. No friend or foe signal received. 1 min till it reaches base area.

Fire Do Not Fire

Better Design

GW/TAI TRUSTWORTHY AI INITIATIVE

Prior Work



Human Only

Human-AI Control

AI Only

Whose action is strictly necessary for 'the-loop' to be complete ?

Both are Required

What is the role of the human(s)?

Act

Human & AI Act

Human-AI Team

Direct AI

AI Acts Only If Human Directs

Is the human presented one or many plans?

One

Human Approver

Many

Human Selector

Human-in-the-loop

Human is Primary

Whose initiative is required for AI to act?

AI

Human Acts AI Oversees

Type of access to the control surface of system?

Direct

AI-on-the-Loop

Indirect

AI-over-the-Loop

Human

Human Acts AI Assists

Type of access to the control surface of system?

Direct

AI-along-the-Loop

Indirect

AI-under-the-Loop

AI is Primary

Whose initiative is required for Human to act?

Human

AI Acts Human Oversees

Type of access to the control surface of system?

Direct

Human-on-the-Loop

When can the human(s) act?

Before AI Action is Final

Command by Veto

While AI Operates

Human Supervisor

Indirect

Human-over-the-Loop

When can the human(s) act?

While AI Operates

Executive Command

After AI Acts

Human Feedback Loop

AI

AI Acts Human Assists

Type of access to the control surface of system?

Direct

Human-along-the-Loop

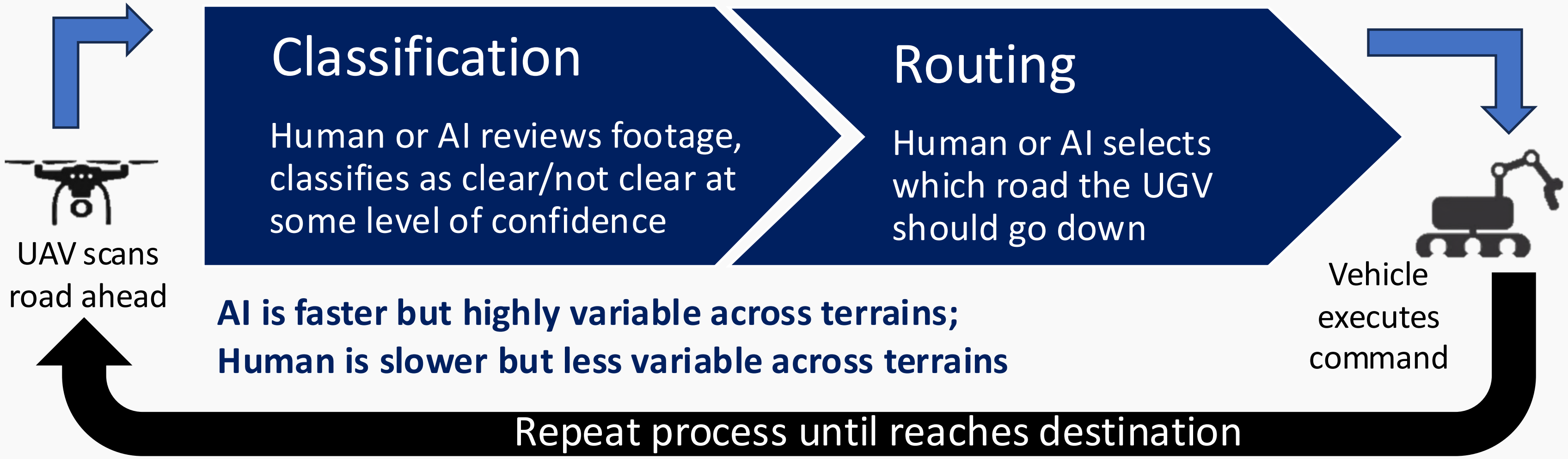
Indirect

Human-under-the-Loop

Research Setting: Minefield Traversal

Using the framework, we modeled several architectures which determined how tasks were allocated between humans and AI and how they worked together

Mine presence may be predicted by sending a UAV to collect data about the road.



UAV scans road ahead

Classification

Human or AI reviews footage, classifies as clear/not clear at some level of confidence

Routing

Human or AI selects which road the UGV should go down

Vehicle executes command

AI is faster but highly variable across terrains; Human is slower but less variable across terrains

Repeat process until reaches destination

Architecture Implementations in Simulation Environment

Classification

Human Approver
Human must approve AI classification, reassign if confidence is too low

Human-AI Team
Classification split up by terrain to take advantage of comparative advantages

AI-on-the-Loop
AI can override human classification if it is more confident in its classification

Human Only (Baseline)
Human does all classification

Both

Human Approver
← →

AI-along-the-Loop
Human can delegate classification / routing decision to AI

AI Only (Baseline)
AI does all classification / routing

Routing

Human Approver
Strategic: AI presents recommended route to execute, human can override for a safer route
Tactical: If a road is classified as unclear, human must approve divergent road choice

Human Selector
Strategic: AI presents possible routes to execute, human selects
Tactical: If a road is classified as unclear, AI presents all possible next roads, human selects

Human-along-the-Loop
AI can delegate decision between roads of similar expected travel time to human

Executive Command
Human can provide AI with guidance on how to operate (conservatively or aggressively)

Command by Veto
Human can override a routing decision but is not required to act

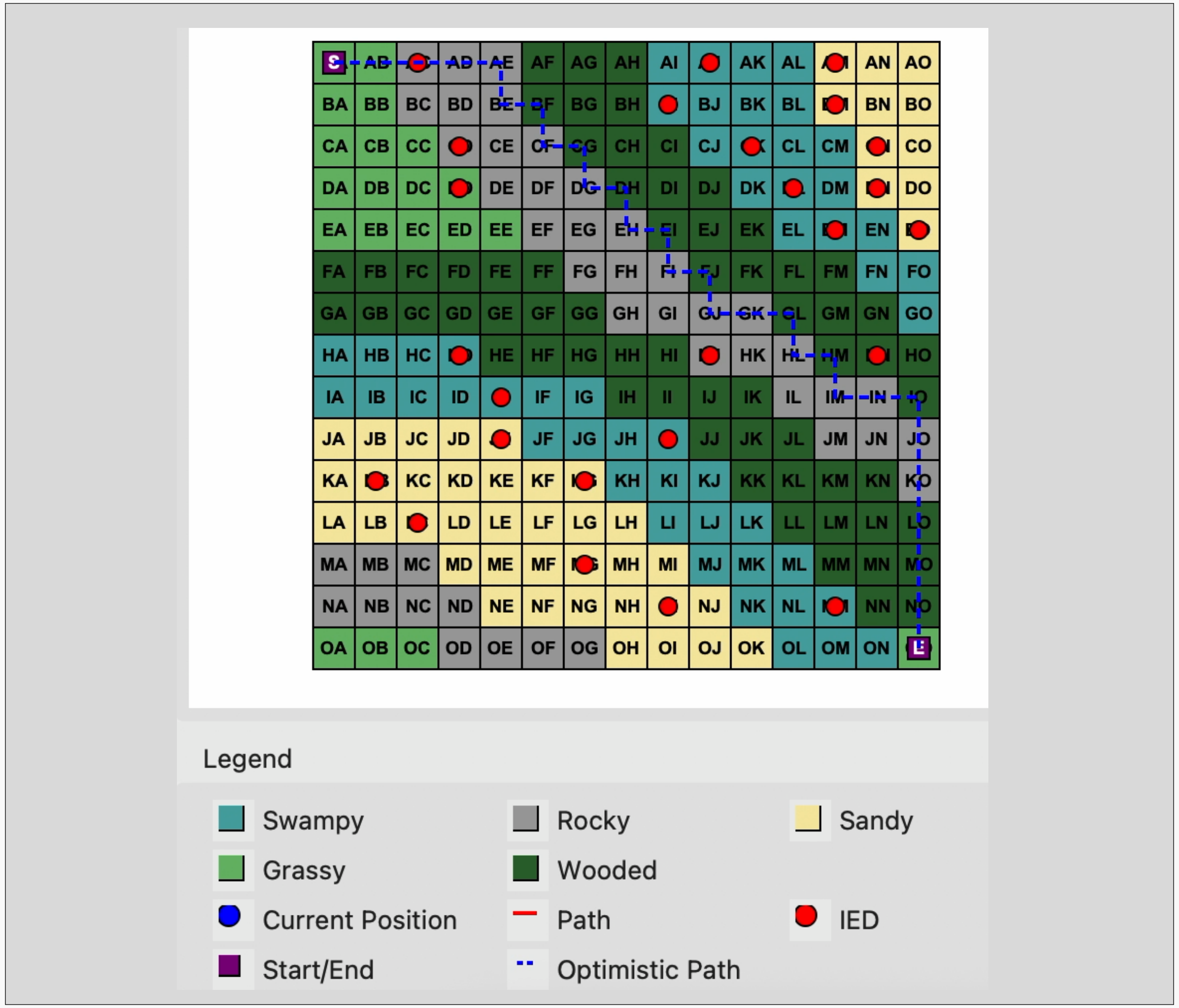
HAI Simulation Set-up

Variables

Treatment:
HAI Architectures

Environment:
Map size
Terrain
IED density
+
Human and AI
confidence

Simulation Testbed



S	AB	●	AD	AE	AF	AG	AH	AI	●	AK	AL	●	AN	AO
BA	BB	BC	BD	BE	BF	BG	BH	●	BJ	BK	BL	●	BN	BO
CA	CB	CC	●	CE	CF	CG	CH	CI	CJ	●	CL	CM	●	CO
DA	DB	DC	●	DE	DF	DG	DH	DI	DJ	DK	●	DM	●	DO
EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	●	EN	●
FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL	FM	FN	FO
GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO
HA	HB	HC	●	HE	HF	HG	HH	HI	●	HK	HL	HM	●	HO
IA	IB	IC	ID	●	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO
JA	JB	JC	JD	●	JF	JG	JH	●	JJ	JK	JL	JM	JN	JO
KA	●	KC	KD	KE	KF	●	KH	KI	KJ	KK	KL	KM	KN	KO
LA	LB	●	LD	LE	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO
MA	MB	MC	MD	ME	MF	●	MH	MI	MJ	MK	ML	MM	MN	MO
NA	NB	NC	ND	NE	NF	NG	NH	●	NJ	NK	NL	●	NN	NO
OA	OB	OC	OD	OE	OF	OG	OH	OI	OJ	OK	OL	OM	ON	E

Legend

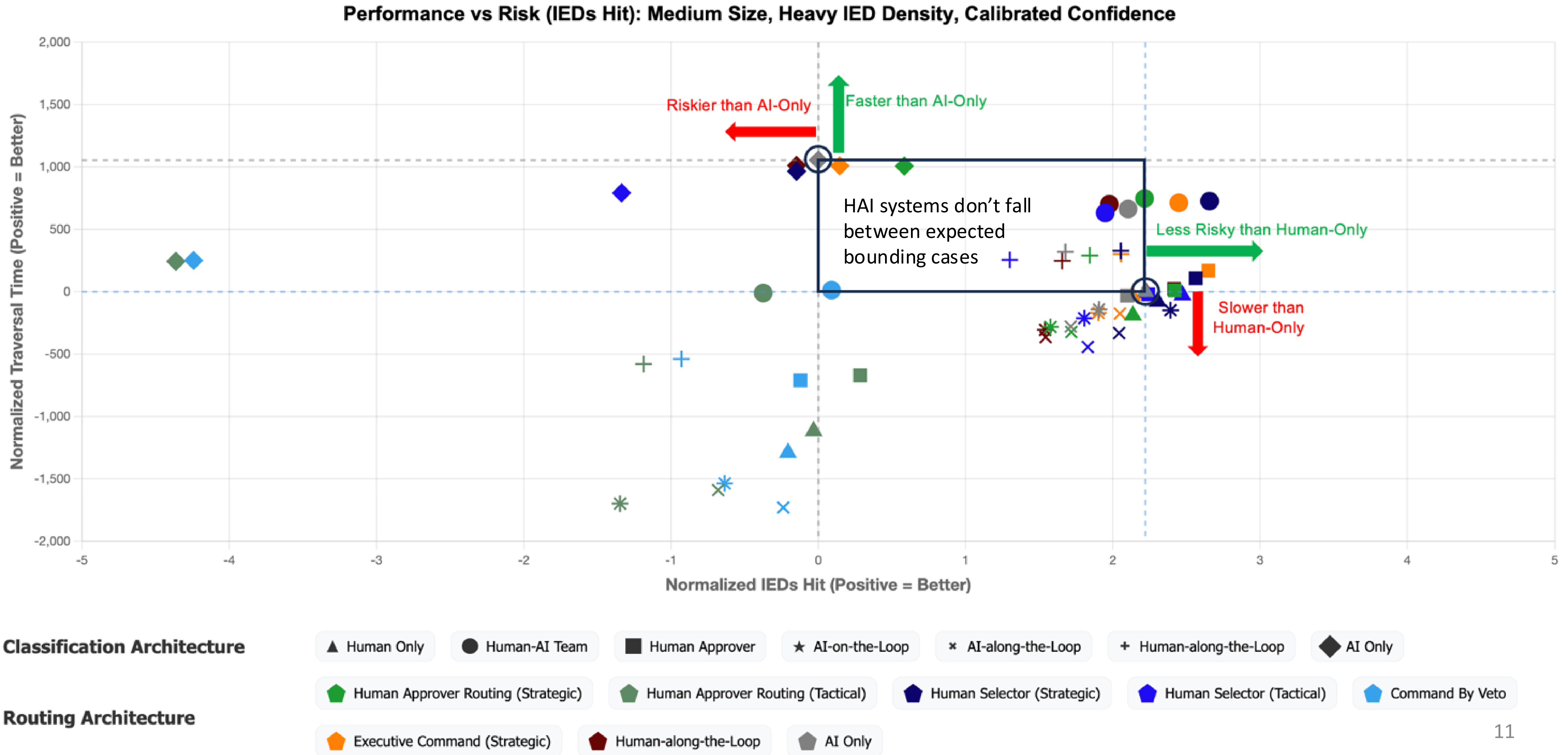
🟦 Swampy	▨ Rocky	🟨 Sandy
🟩 Grassy	■ Wooded	
● Current Position	— Path	● IED
■ Start/End	-- Optimistic Path	

FOMs

Performance:
Traversal time
(normalized)
Path length

Risk:
IEDs hit
(normalized)
Classification
errors
(type 1 and 2)

HAI's Non-Linear Tradeoffs



Performance vs Risk (IEDs Hit): Medium Size, Heavy IED Density, Calibrated Confidence

Riskier than AI-Only

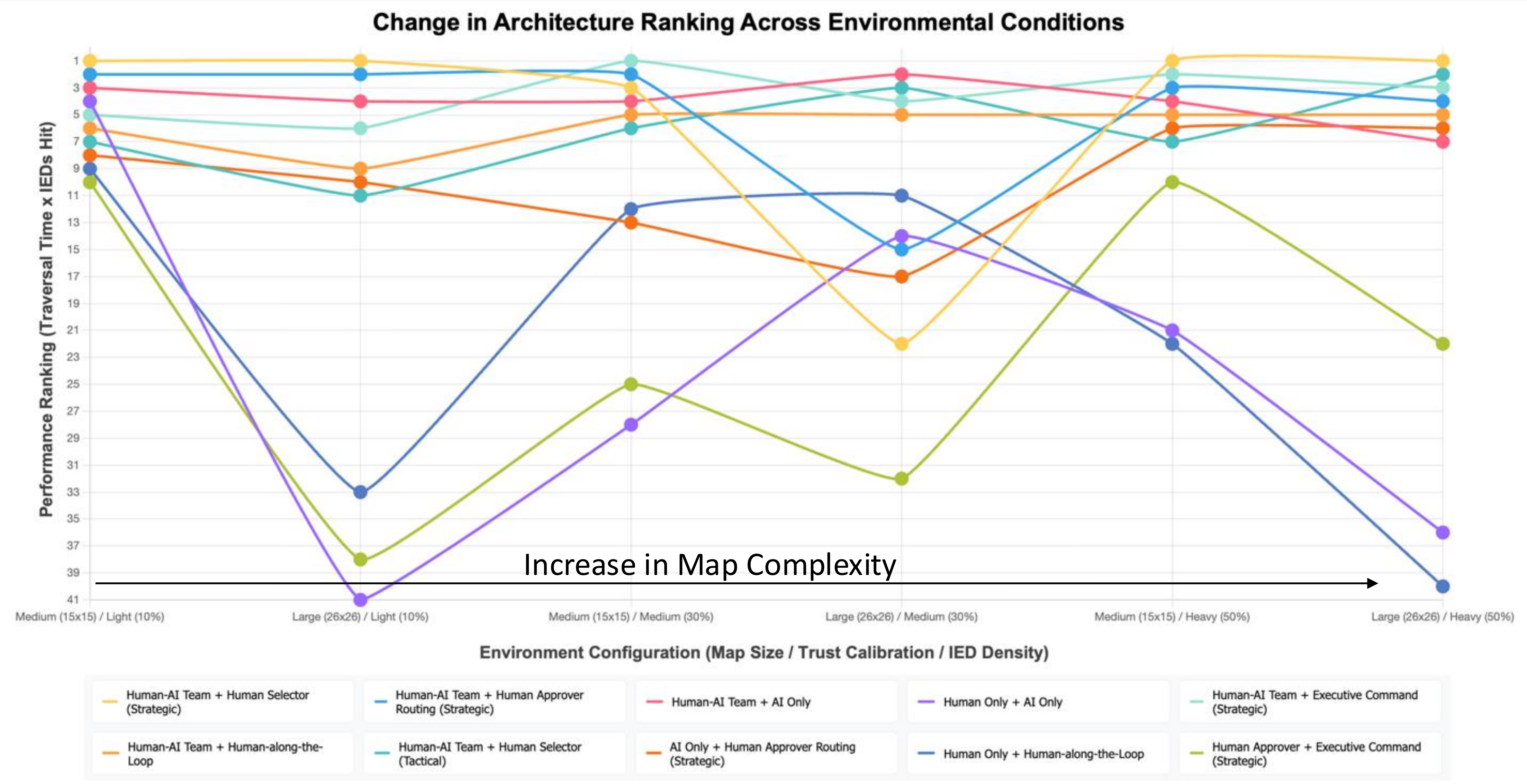
Faster than AI-Only

HAI systems don't fall between expected bounding cases

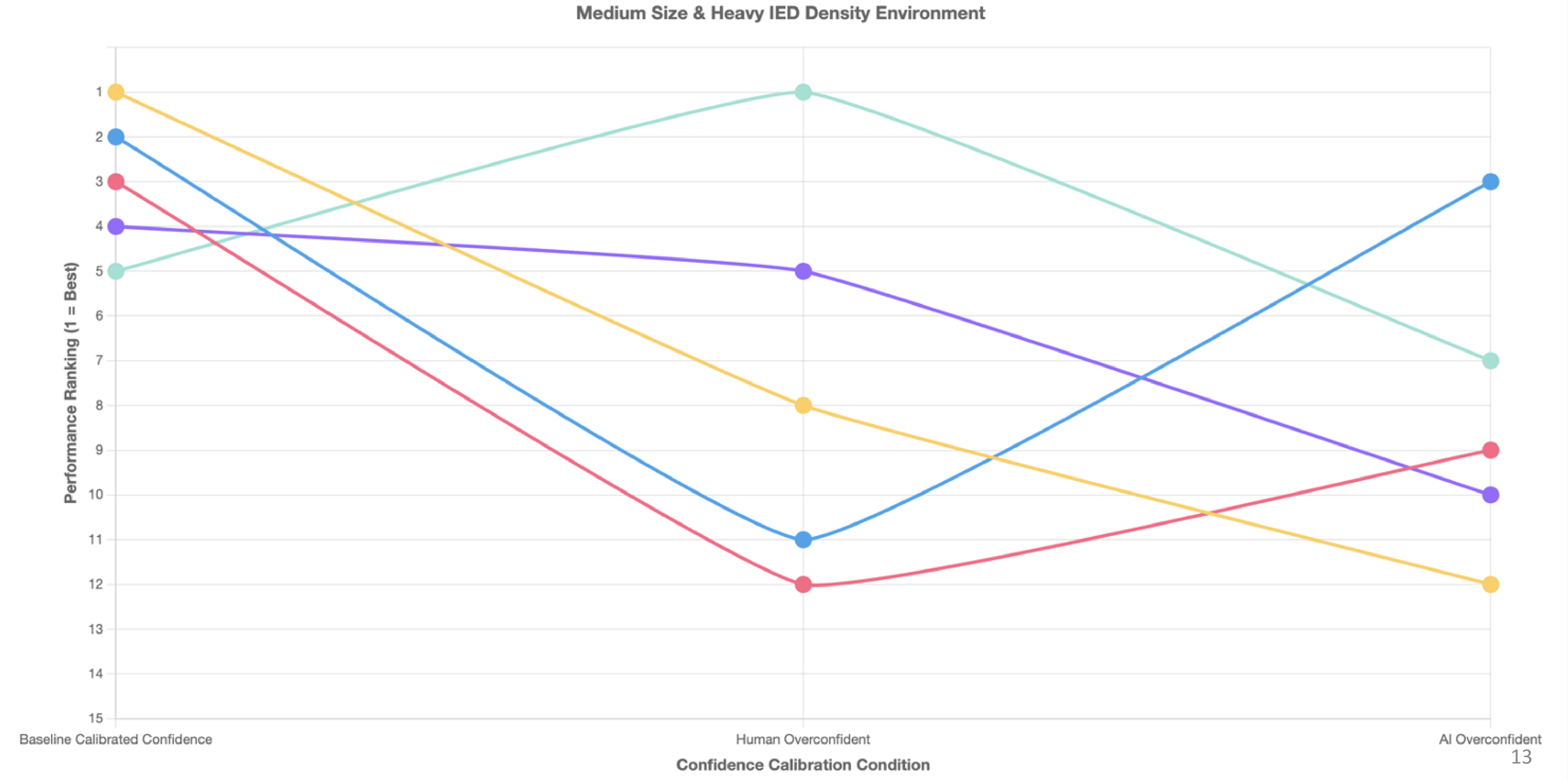
Less Risky than Human-Only

Slower than Human-Only

Robustness of Results to Context



Importance of Training on Interaction



Findings

- **Architecture, Environment, and their Interaction are all significant**
  - All three were statistically significant in ANOVA tests
- **Change in performance across environmental conditions was not uniform, consistent, or obvious**
  - Seemingly innocuous changes in operating environment (increasing map size with same IED density and confidence) led to large changes in relative performance for some architectures

Implications for Test & Evaluation

- Need to expand system boundary of T&E to consider human-AI architecture & interaction
  - Changing just *how* the human is integrated significantly changed results while holding the technical performance constant
- Human-AI systems testbeds can:
  - Reveal non-obvious tradeoffs and interactions
  - Understand how changing variables affect system outcomes
  - Identify which architectures that are robust / sensitive to expected operating environment

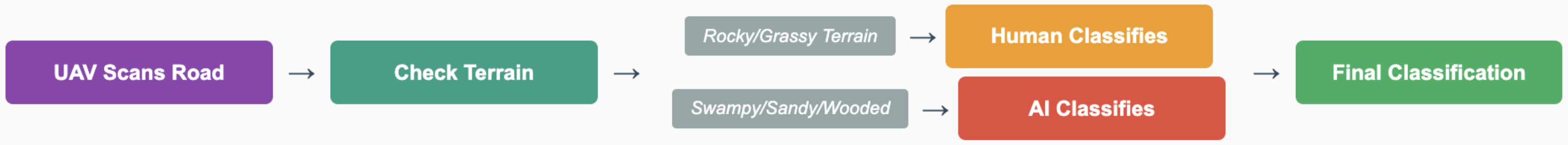
Thank You

asingh25@gwu.edu

Classification Architectures

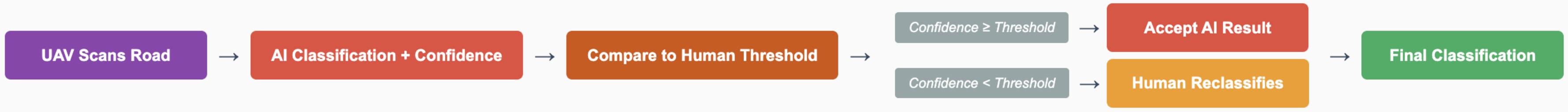
Human-AI Team

Leverages complementary strengths based on historical performance



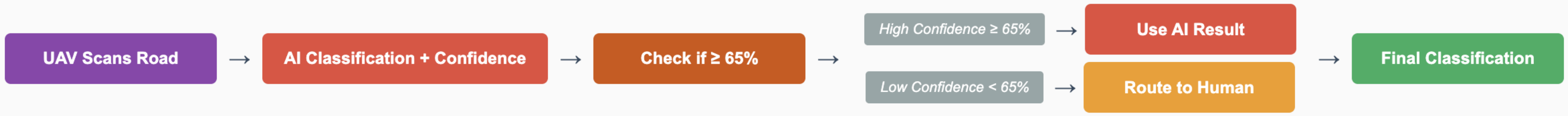
Human Approver

Threshold adjusts dynamically: decreases per correct AI decision (building trust), increases 20% after IED encounter (betrayal).



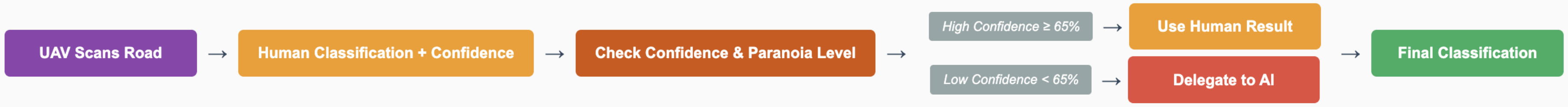
Human-along-the-Loop

Simple rules-based system with fixed threshold, reassigns low-confidence cases to human expert.



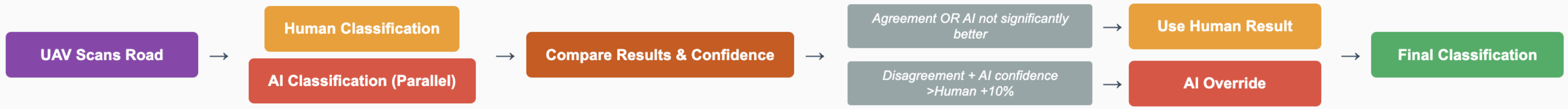
AI-along-the-Loop

Models paranoia that increases with consecutive "clear" classifications. Paranoia resets when mines are found or roads marked unclear.



AI-on-the-Loop

AI monitors and only overrides when it disagrees AND has significantly higher confidence (+10% margin).



UAV Scans Road
Human Classification
AI Classification (Parallel)
Compare Results & Confidence
Agreement OR AI not significantly better
Use Human Result
Disagreement + AI confidence >Human +10%
AI Override
Final Classification

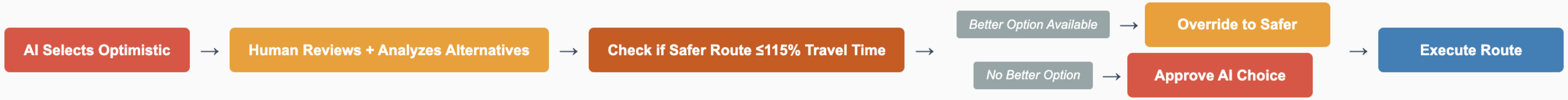
GW/TAI TRUSTWORTHY AI INITIATIVE

Routing Architectures (Strategic vs Tactical)

Human makes decisions about what type of route should be executed

Human Approver (Strategic)

Human can substitute a safer path (more favorable terrain) if expected travel time is within 15% of shortest path.



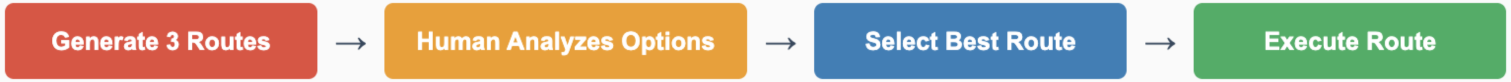
AI Selects Optimistic → Human Reviews + Analyzes Alternatives → Check if Safer Route ≤115% Travel Time →

Better Option Available → Override to Safer → Execute Route

No Better Option → Approve AI Choice → Execute Route

Human Selector (Strategic)

Route Options: 1) Shortest path, 2) AI-favorable, 3) Human-favorable.
Selection Logic: Human-AI Team → shortest path. AI-dominant systems → AI-favorable. Human-dominant systems → Human-favorable.

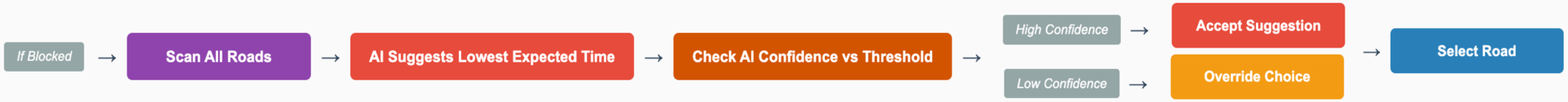


Generate 3 Routes → Human Analyzes Options → Select Best Route → Execute Route

Human makes decisions when an issue occurs

Human Approver (Tactical)

If confidence is low, human selects next appropriate road classified as clear with highest confidence.



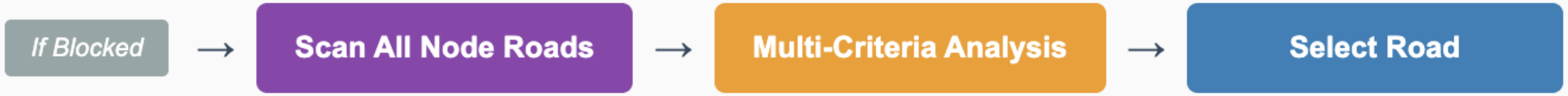
If Blocked → Scan All Roads → AI Suggests Lowest Expected Time → Check AI Confidence vs Threshold →

High Confidence → Accept Suggestion → Select Road

Low Confidence → Override Choice → Select Road

Human Selector (Tactical)

Human weighs expected travel time, progress toward goal, and AI classification performance in different terrains.



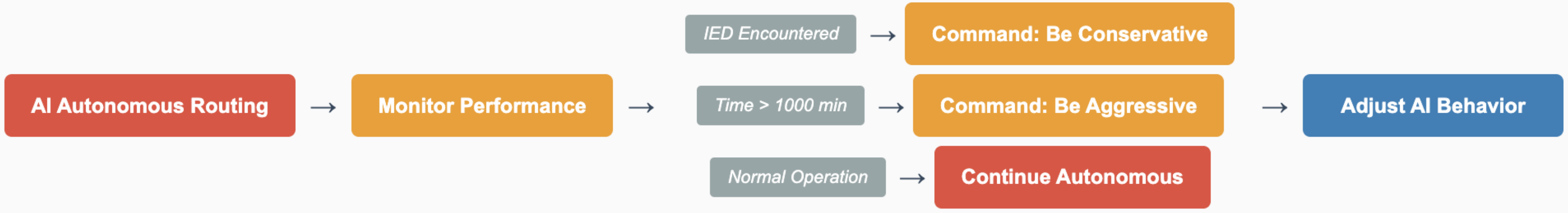
If Blocked → Scan All Node Roads → Multi-Criteria Analysis → Select Road

GW TAI
TRUSTWORTHY
AI INITIATIVE

Routing Architectures (Only One Implementation)

Executive Command

Human adjusts AI behavior. Conservative mode prioritizes roads where AI has high classification confidence over pure expected value. Aggressive mode emphasizes progress toward end node over safety margins


AI Autonomous Routing → Monitor Performance →

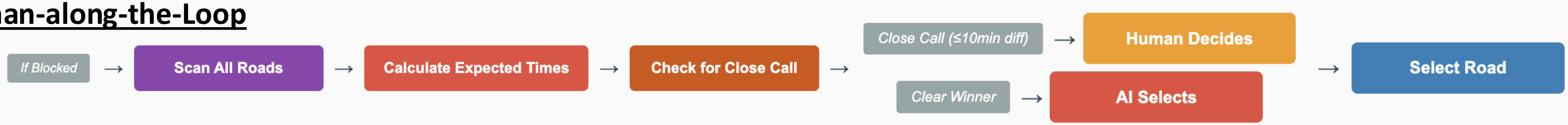
IED Encountered → Command: Be Conservative

Time > 1000 min → Command: Be Aggressive

Normal Operation → Continue Autonomous

→ Adjust AI Behavior

Human-along-the-Loop


If Blocked → Scan All Roads → Calculate Expected Times → Check for Close Call →

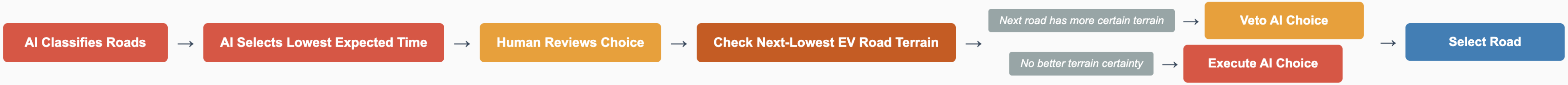
Close Call (≤10min diff) → Human Decides

Clear Winner → AI Selects

→ Select Road

When two or more road options have expected travel times within 10 minutes, AI delegates to human expertise. Human Selects the road classified as clear with highest confidence level that makes progress toward destination.

Command by Veto


AI Classifies Roads → AI Selects Lowest Expected Time → Human Reviews Choice → Check Next-Lowest EV Road Terrain →

Next road has more certain terrain → Veto AI Choice

No better terrain certainty → Execute AI Choice

→ Select Road

Human can reject AI's lowest expected travel time choice if the next-lowest EV road has more certain terrain type for AI classification but is not required to act for AI to operate