Understanding the Tradeoffs of Human-AI System Architecting

Aditya Singh, Zoe Szajnfarber

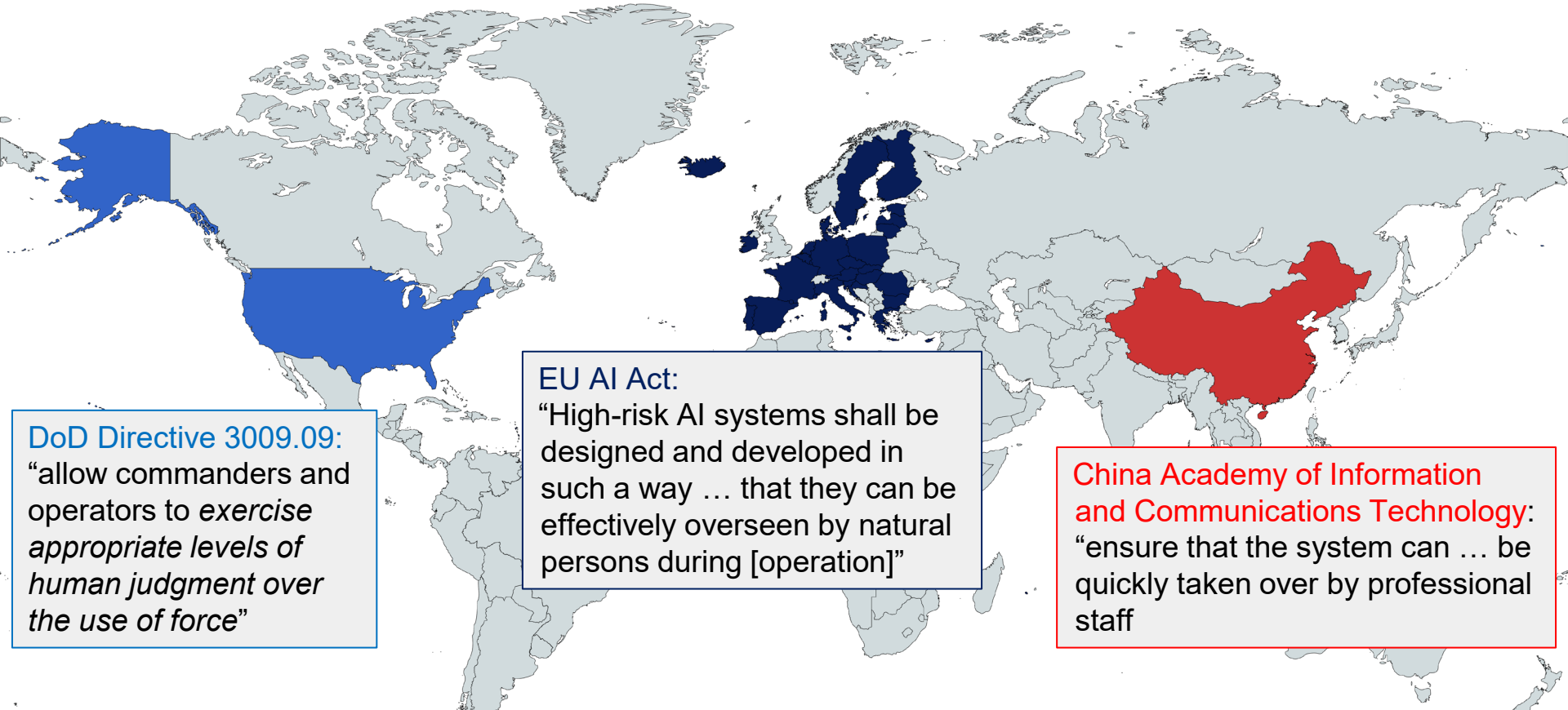
GW TAI
TRUSTWORTHY AI INITIATIVE

Human Oversight: The Silver Bullet for Trustworthy AI?

"One commonly proposed principle among researchers and the military alike is that there should be a 'human in the loop' of autonomous weapons. But where and how people should or must be involved is still up for debate."

D. Adam, "Lethal AI weapons are here: how can we control them?," *Nature*, vol. 629, no. 8012, pp. 521–523, Apr. 2024, doi: 10.1038/d41586-024-01029-0.

Human Control: A New Policy Prescription



DoD Directive 3009.09:
"allow commanders and operators to *exercise appropriate levels of human judgment over the use of force*"

EU AI Act:
"High-risk AI systems shall be designed and developed in such a way … that they can be effectively overseen by natural persons during [operation]"

China Academy of Information and Communications Technology:
"ensure that the system can … be quickly taken over by professional staff

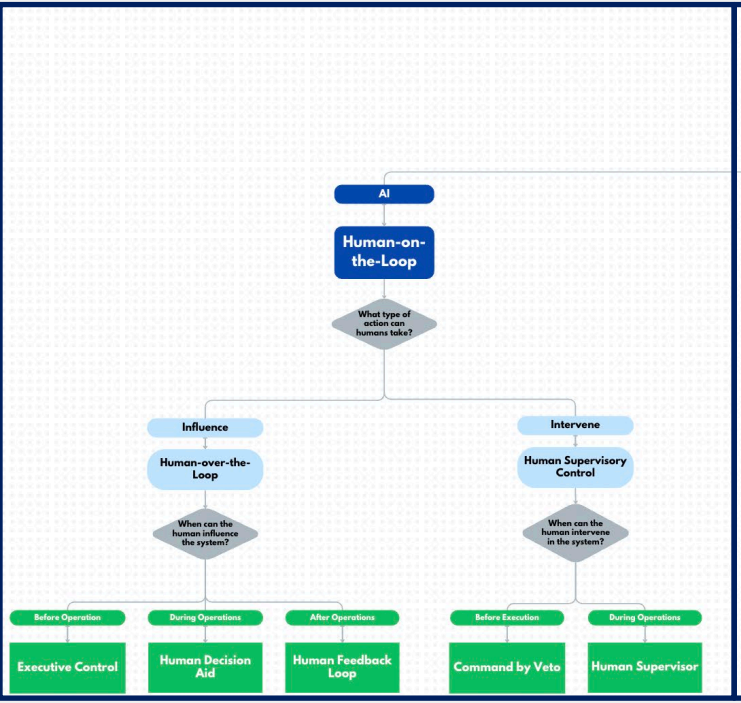
Lack of Clarity on ”Human Control”

	Human-in-the-Loop	Human-on-the-Loop
DoD	“only engage individual targets or specific target groups that have been selected by a human operator”	“operators have the ability to monitor and halt a weapon’s target engagement”

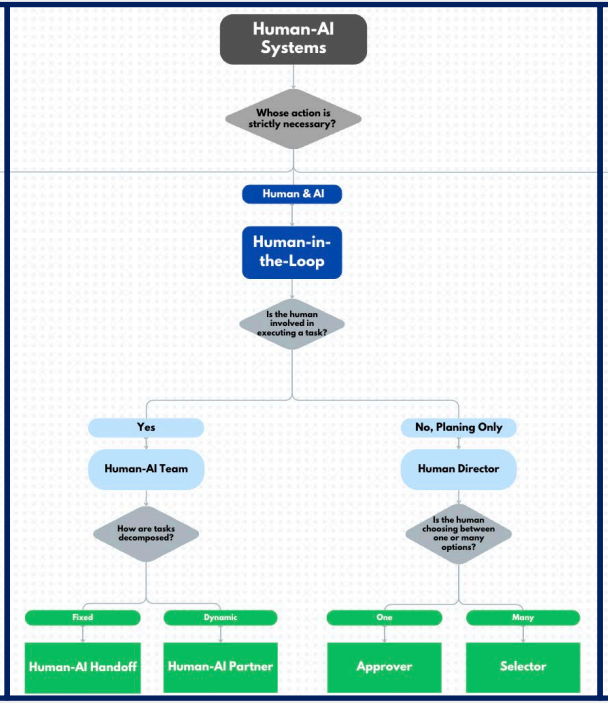
High level definitions mask the complexity of how humans and AI can be partnered together

Prior Work: Defining Human-AI Systems

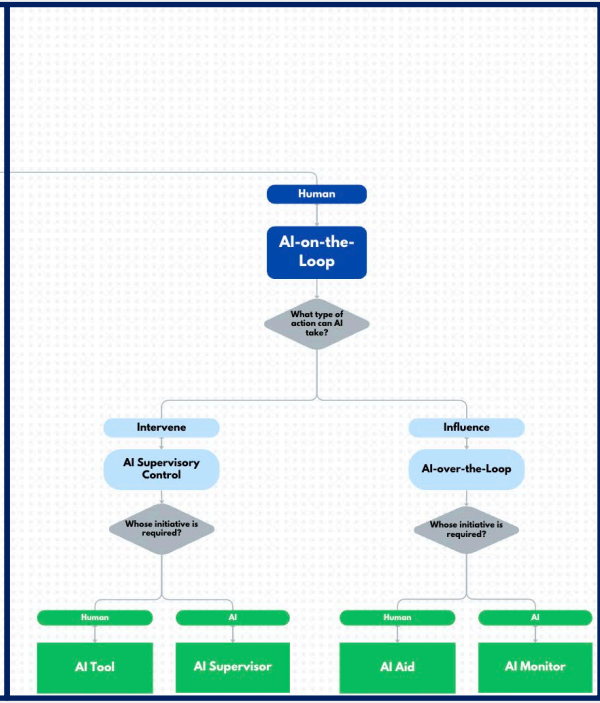
Human Supervising AI



Human-AI Teaming

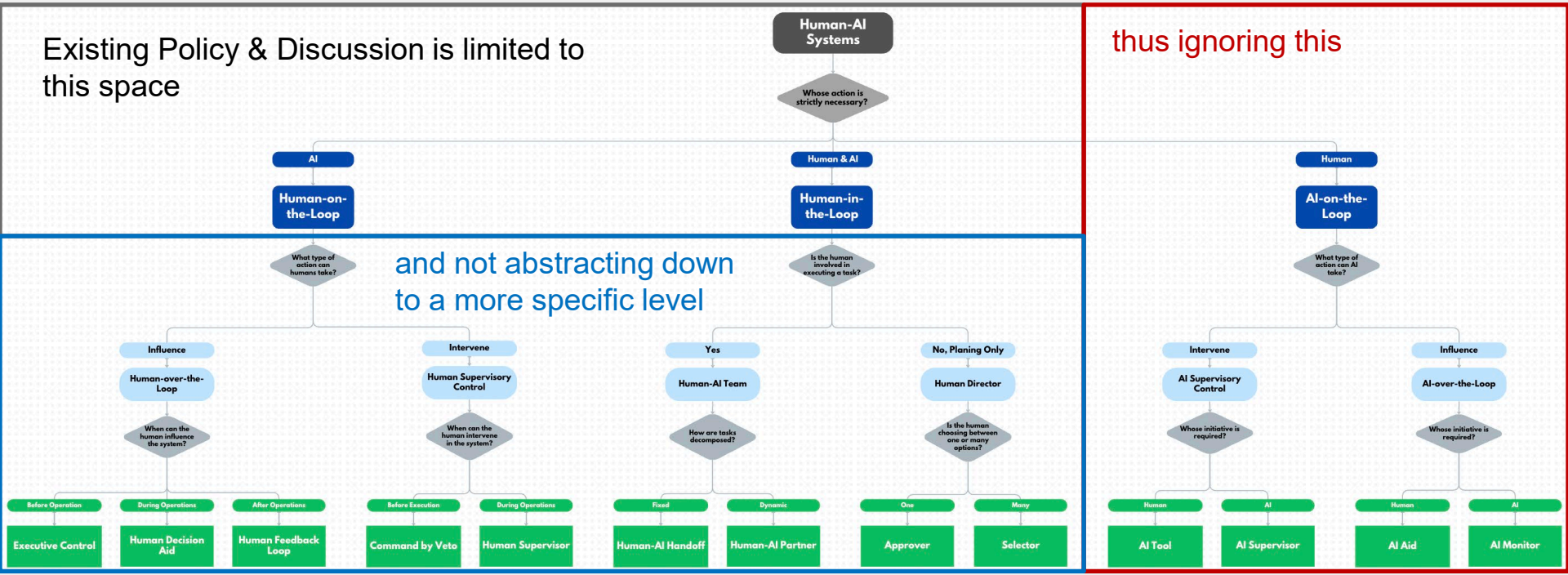


AI Supervising Humans



Singh, Aditya and Szajnfarber, Zoe, Preposition Salad: Making Sense of Human-in/on/over-the-Loop Control for AI Systems (August 07, 2024). Available at SSRN: https://ssrn.com/abstract=4921359

Prior Work: Why Existing Definitions Fall Short



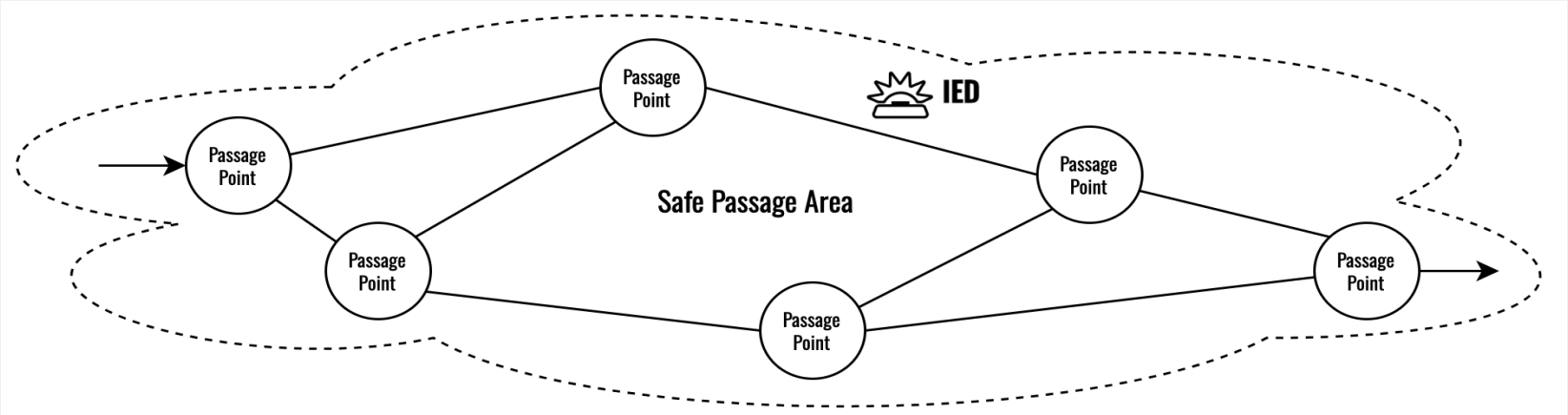
Expanded two high-level concepts to 11 specific architectures

Singh, Aditya and Szajnfarber, Zoe, Preposition Salad: Making Sense of Human-in/on/over-the-Loop Control for AI Systems (August 07, 2024). Available at SSRN: https://ssrn.com/abstract=4921359

Research Goal: Understand Tradeoffs

Apply these architectures to a common reference problem to understand the tradeoffs associated with each

Silverfish Problem



- Mission performance defined as time to clear a path from start to end
- Understand how to design AI into a notional system and characterize the risk vs performance tradeoffs of doing so

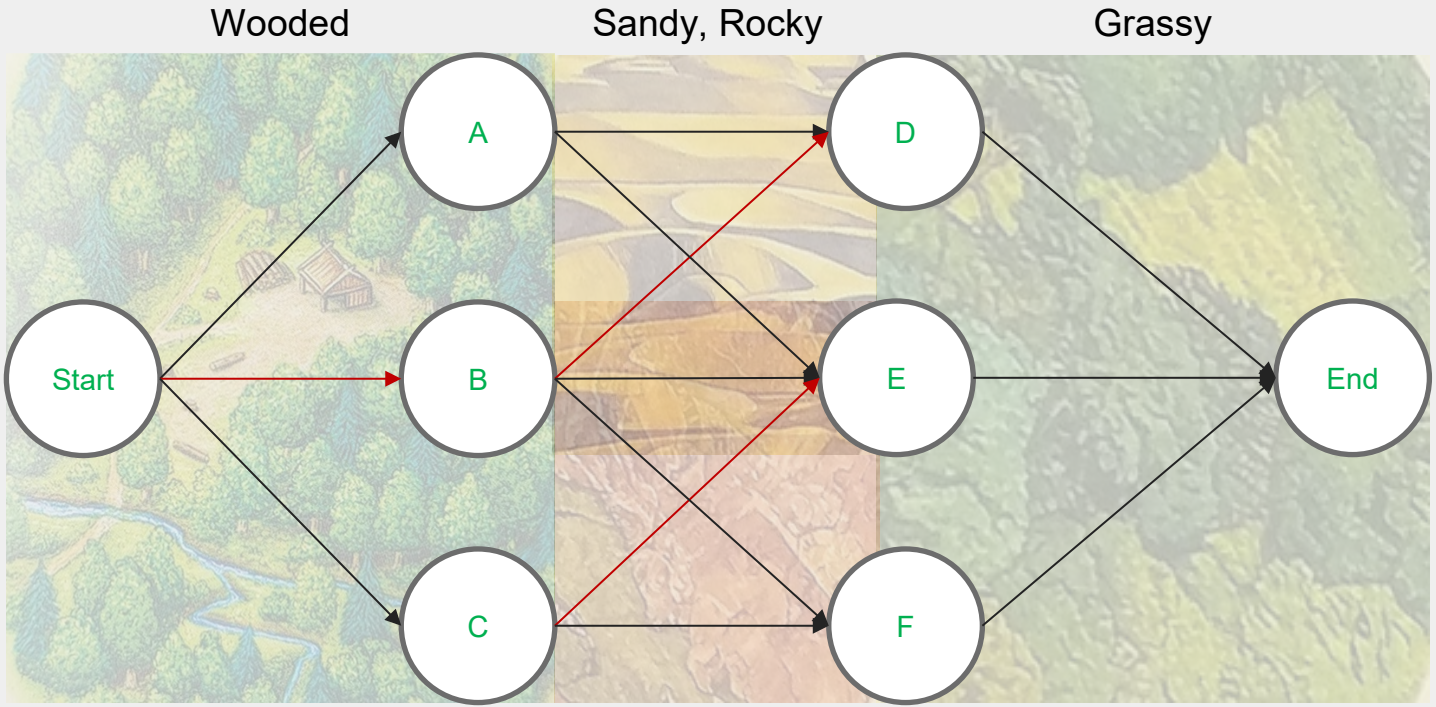
Silverfish Key Resources

UAV takes 1 minute per link to scan and report data back

UGV guides troops through each link in 20 minutes. If a mine is on the link, the link takes an additional 40 minutes to clear.

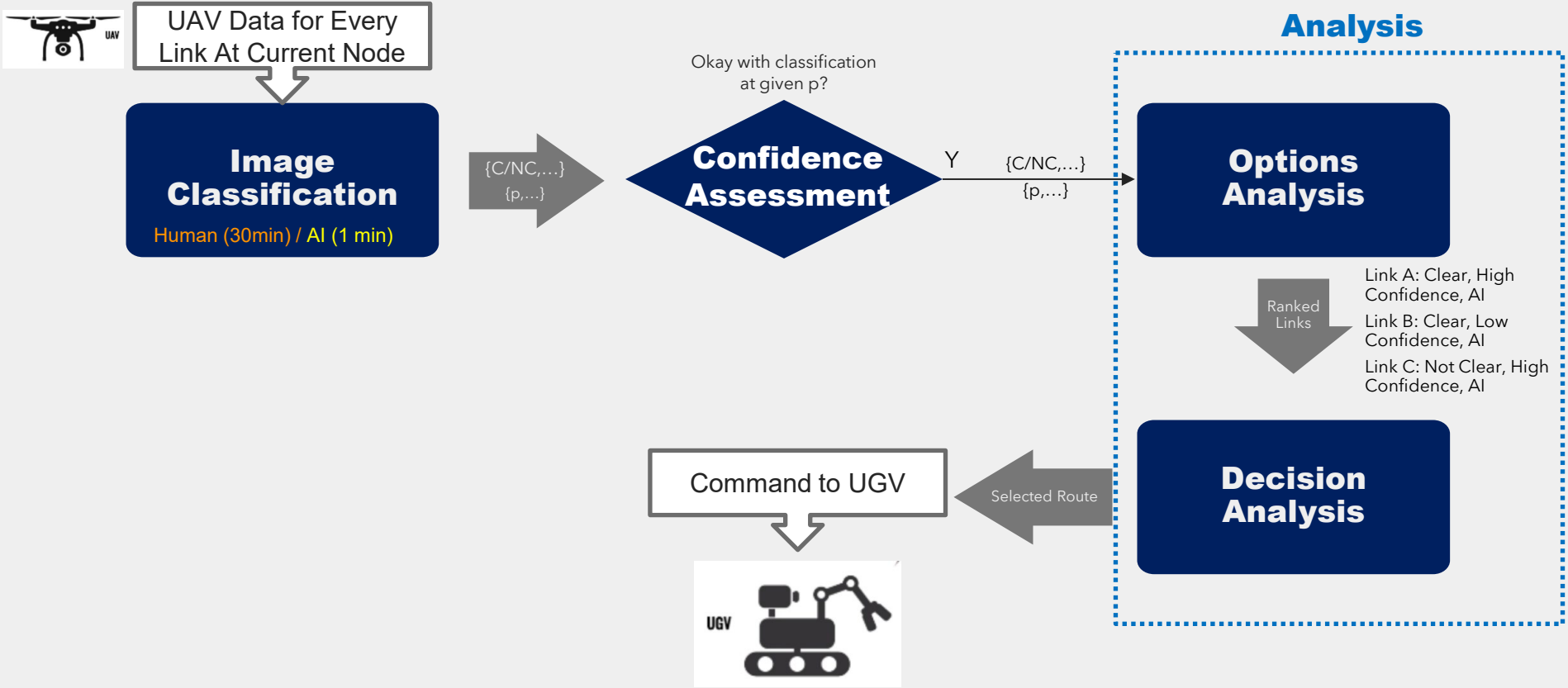
AI system estimate likelihood of a mine per link in 1 minute. Alternatively, a human expert can analyze the link but takes 30 minuets. AI performance is highly variable, while human expert has less variance in their accuracy.

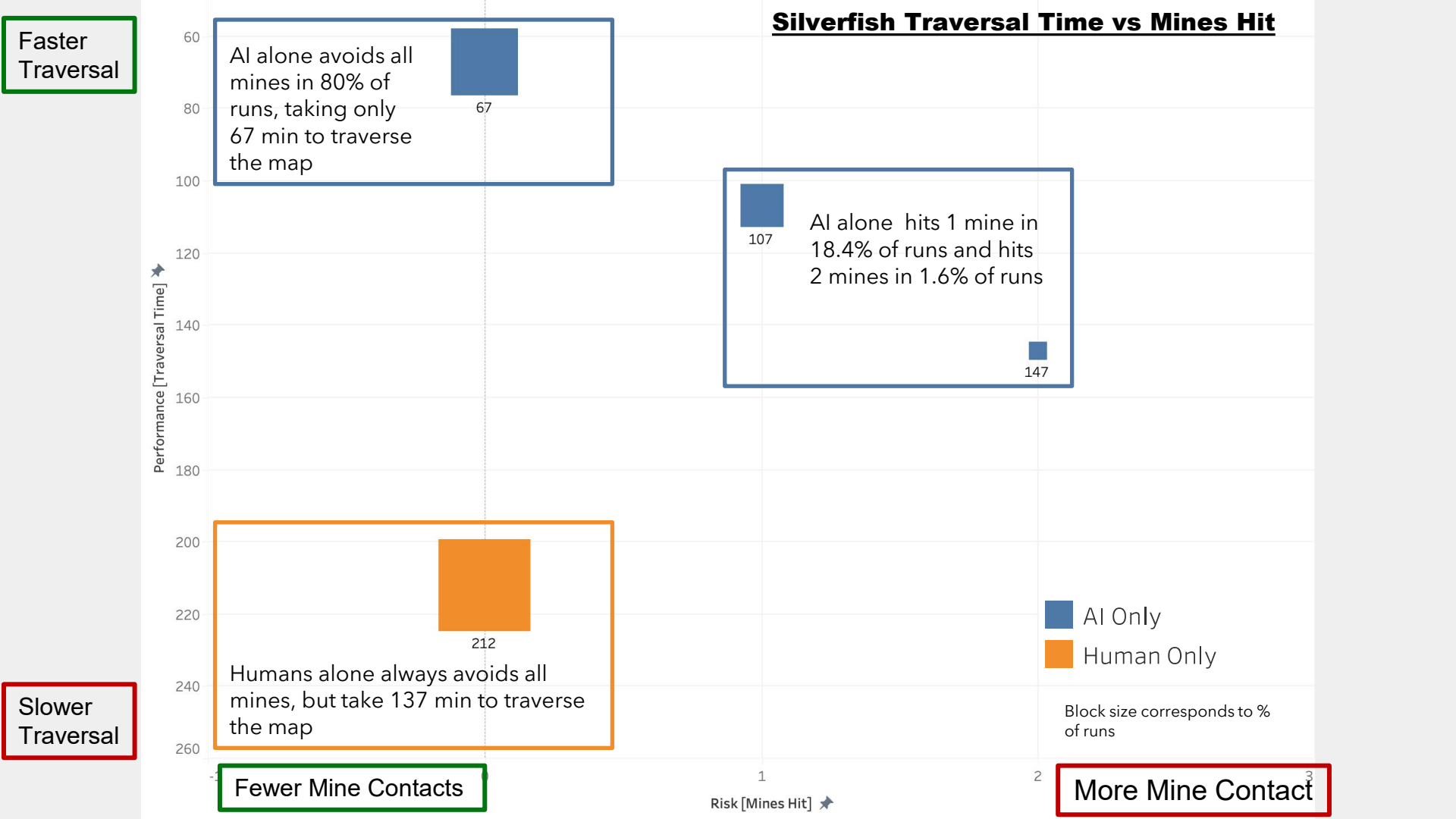
SilverFish Map



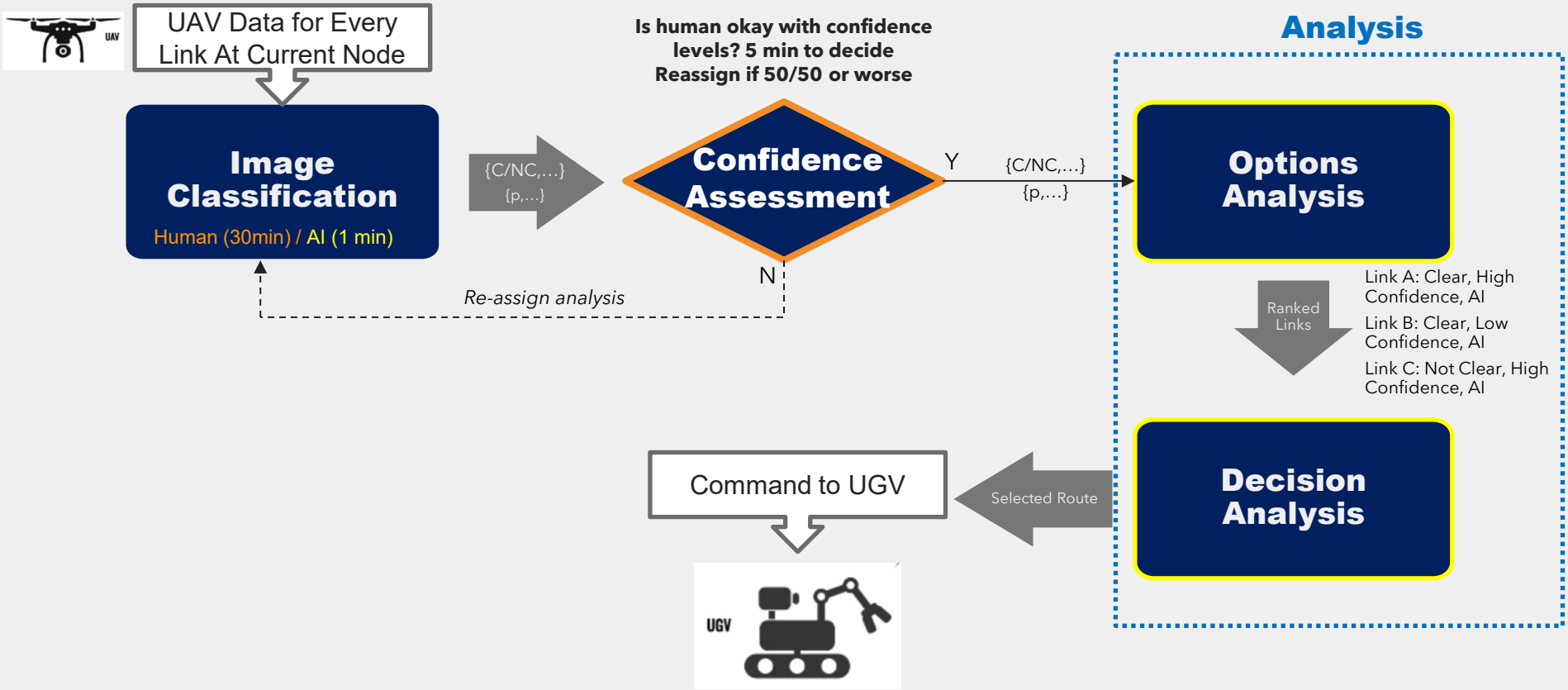
Accuracy is affected by environmental conditions of links

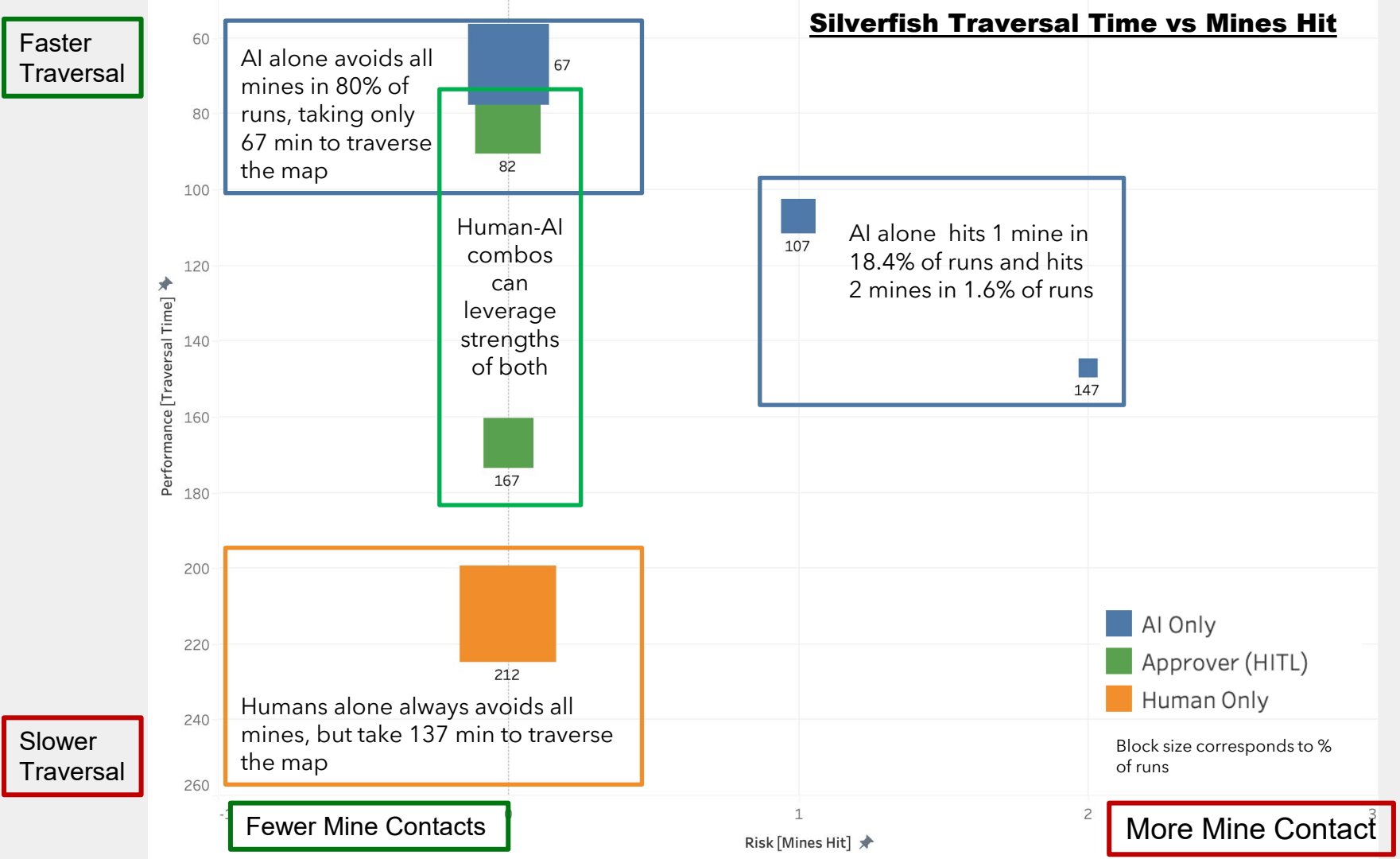
Simplified Decision Flow




Faster Traversal
Slower Traversal
Silverfish Traversal Time vs Mines Hit
AI alone avoids all mines in 80% of runs, taking only 67 min to traverse the map
67
AI alone hits 1 mine in 18.4% of runs and hits 2 mines in 1.6% of runs
107
147
Humans alone always avoids all mines, but take 137 min to traverse the map
212
AI Only
Human Only
Block size corresponds to % of runs
Performance [Traversal Time]
60
80
100
120
140
160
180
200
220
240
260
Risk [Mines Hit]
1
2
3
Fewer Mine Contacts
More Mine Contact

Human Approver




Silverfish Traversal Time vs Mines Hit
Faster Traversal
Slower Traversal
AI alone avoids all mines in 80% of runs, taking only 67 min to traverse the map
67
82
Human-AI combos can leverage strengths of both
167
AI alone hits 1 mine in 18.4% of runs and hits 2 mines in 1.6% of runs
107
147
Humans alone always avoids all mines, but take 137 min to traverse the map
212
AI Only
Approver (HITL)
Human Only
Block size corresponds to % of runs
Performance [Traversal Time]
60
80
100
120
140
160
180
200
220
240
260
Risk [Mines Hit]
1
2
3
Fewer Mine Contacts
More Mine Contact

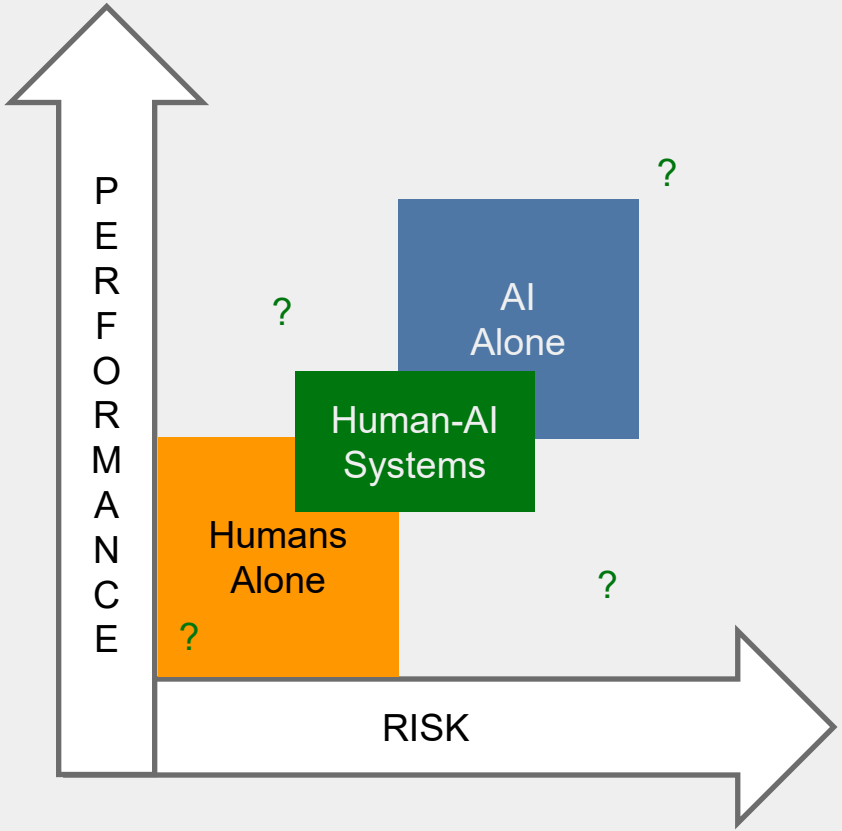
Takeaways

	AI Alone	Human-AI Collaboration	Humans Alone
Avg Time	75.6 min	101.8 min	212 min
Avg Mines Hit	0.26	0.0	0.0

AI performance is superior but at the cost of higher risk; inverse for humans

Human-AI collaboration can leverage AI performance with human judgement

Implications & Future Work



Tradeoffs are not linear

Some architectures may not provide a clear advantage

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

asingh25@gwu.edu

DTAIS

GW Co-Design of Trustworthy AI Systems

TRAILS

Trustworthy AI in Law & Society