



A Novel Framework for Identifying the Threshold of Trust-worthiness of Artificial Intelligence

Dr Tim Forsyth, Ajay Kumar, Dr John Economou

12/Oct/2021

www.cranfield.ac.uk

Section #1

Section #2

Section #3

Trustworthiness and AI

(An Account of Trustworthiness fit for Purpose)

Trustworthiness Framework

(Technical Specification)

Enactment of Framework

(Fuzzy Logic)

- Background:

- Requirement for AI

- Ethical Concerns Limiting AI

- Methodologies concerning the trustworthiness are not fit for purpose.

- The Defence Command Plan (2021) identifies a clear need for AI autonomous capabilities spanning the five domains as being essential to defence modernisation.
- However, this need is hampered by the need for public debate over the moral, legal and ethical concerns of deploying military AI autonomous systems (MoD, 2018).
- There is a failure to adequately assess AI systems trustworthiness (Stanton and Jensen, 2021).



- **What is Trust?**

“Trust is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action”

(Gambetta, 2000)

Only when we understand the point of the concept of trust that supports
‘human-AI-teaming’
can we build an account of trust that is fit for purpose.

Gambetta, Diego (2000) ‘Can We Trust Trust?’, in Gambetta, Diego (ed.) Trust: Making and Breaking Cooperative Relations, electronic edition, Department of Sociology, University of Oxford, chapter 13, pp. 213-237

A Foundation of Trust – Military Necessity

Luis Moreno-Ocampo (2006, p5), Chief Prosecutor of the International Criminal Court, writes:

“The application of Article 8(2)(b)(iv) requires, inter alia, an assessment of:

- a) the anticipated civilian damage or injury;
- b) the anticipated military advantage; and
- c) whether (a) was “clearly excessive” in relation to (b).”

How does the commander/leader task an AI (under conditions of uncertainty) and assess whether the AI is worthy of their trust and the dependency they need to place on the AI?



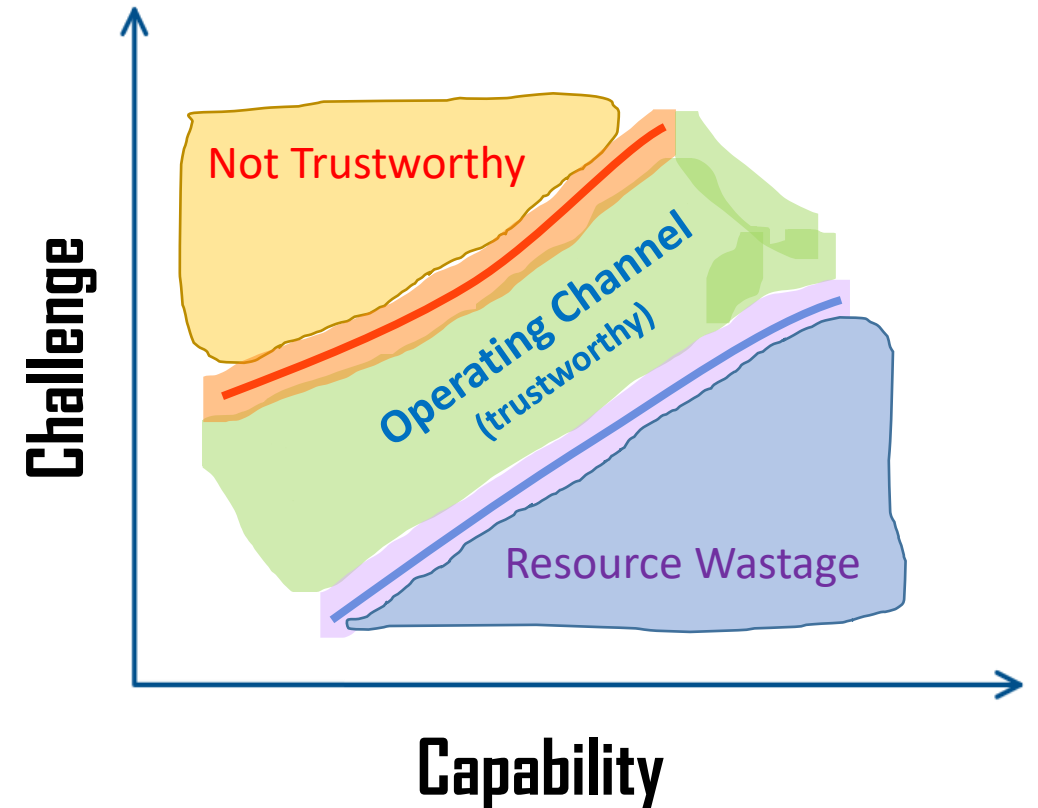
Benefits of the Trustworthiness Framework

- The framework renders a transparent relationship between the AI's capability and suitability to the challenge it is tasked. This approach paves the way for debate on ethical and legal concerns.
- The framework provides a structure that can be directly applied across all five domains of defence (space, air, land, sea and cyber).
- It is not necessary to have a technical background to understand and interpret the framework.
- The variables of the framework can be combined using technical (transparent and traceable) means to capture the complexities of the nuances between the variables into a coherent picture of trustworthiness.

Assessing The Trustworthiness of AI

The identification of the AI **suitability** to the task is achieved by dividing the operational space of the AI into three explicit domains:

- Optimal performance (trustworthy) – a balance between AI capability and mission challenge.
- Inefficiently tasked – AI capability leads mission challenge.
- **NOT** Trustworthy – Challenge is overwhelming the AI capability.



Stamp, G., (1989). The Individual, the Organisation and the Path to Mutual Appreciation. *Personnel Management*, 21(7), pp28-31

Section #1

Section #2

Section #3

Trustworthiness and AI

(An Account of Trustworthiness fit for Purpose)

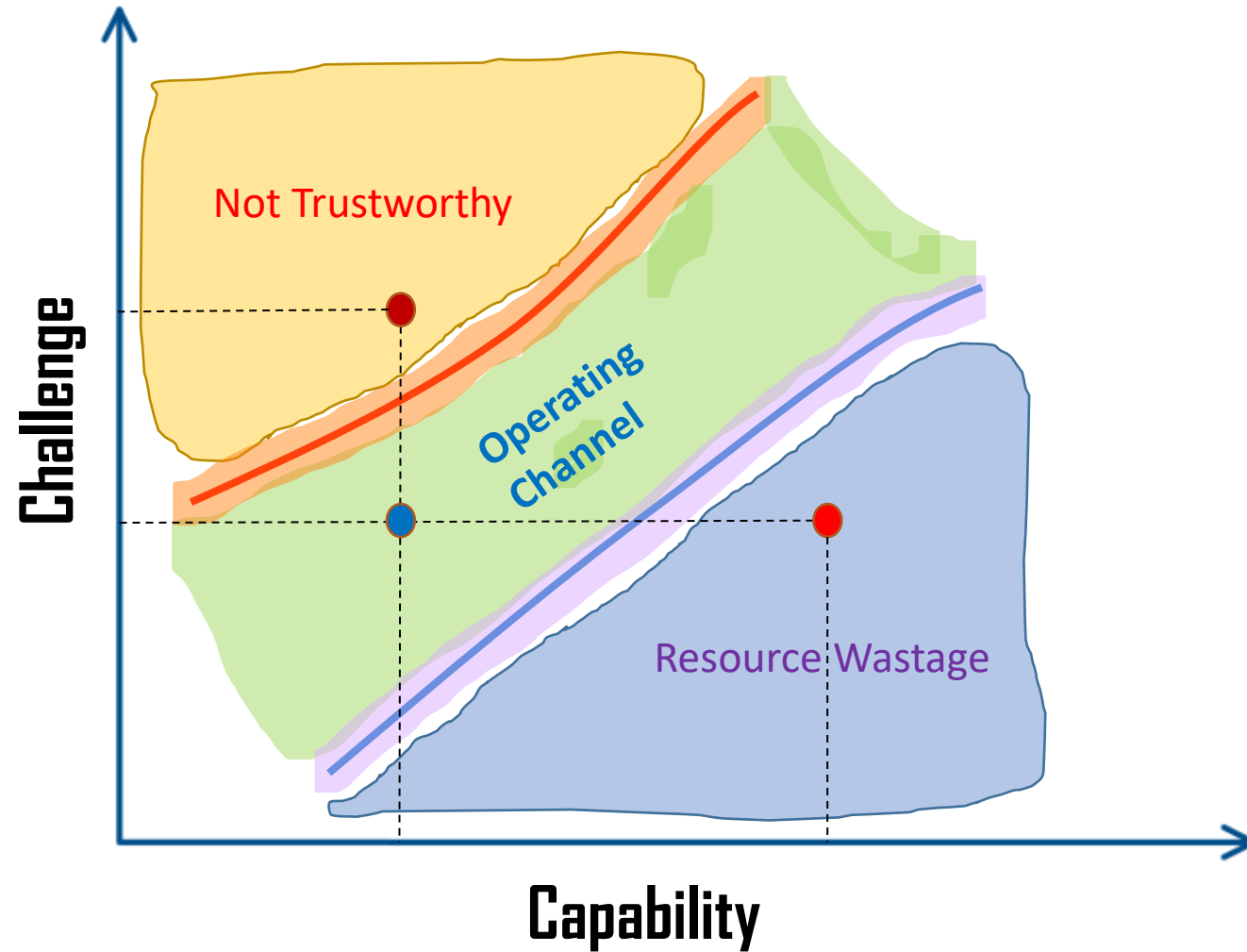
Trustworthiness Framework

(Technical Specification)

Enactment of Framework

(Fuzzy Logic)

Framework



$$Capability = K \times (Means)^p \times (Ways)^q \times (Suitability)^r$$

Auto Encoder (AE)
Boltzman Machine (BM)
Convolutional Neural Net (CNN)
Decision Trees (DT)
Deconvolutional Network (DN)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Deep Feed Forward (DFF)
Deep Q-networks
Deep Residual Network (DRN)
Denoising Auto Encoder (DAE)
Echo State Network (ESN)
Extreme Learning Machine (ELM)
Feed Forward (FF)
Gated Recurrent Unit (GRU)
Generative Adversarial Network (GAN)
Hopfield Network (HN)
Kohonen Network (KN)

Liquid State Machine (LSM)
Long/Short Term Memory (LSTM)
Markov Chain (MC)
Neural Turing Machine (NTM)
Perceptron
Radial Basis Network (RBF)
Random Forest (RF)
Recurrent Neural Network (RNN)
Restricted Boltzmann Machine (RBM)
Sparse Auto Encoder (SAE)
Support Vector Machine (SVM)
Variational Auto Encoder (VAE)
... and so on.

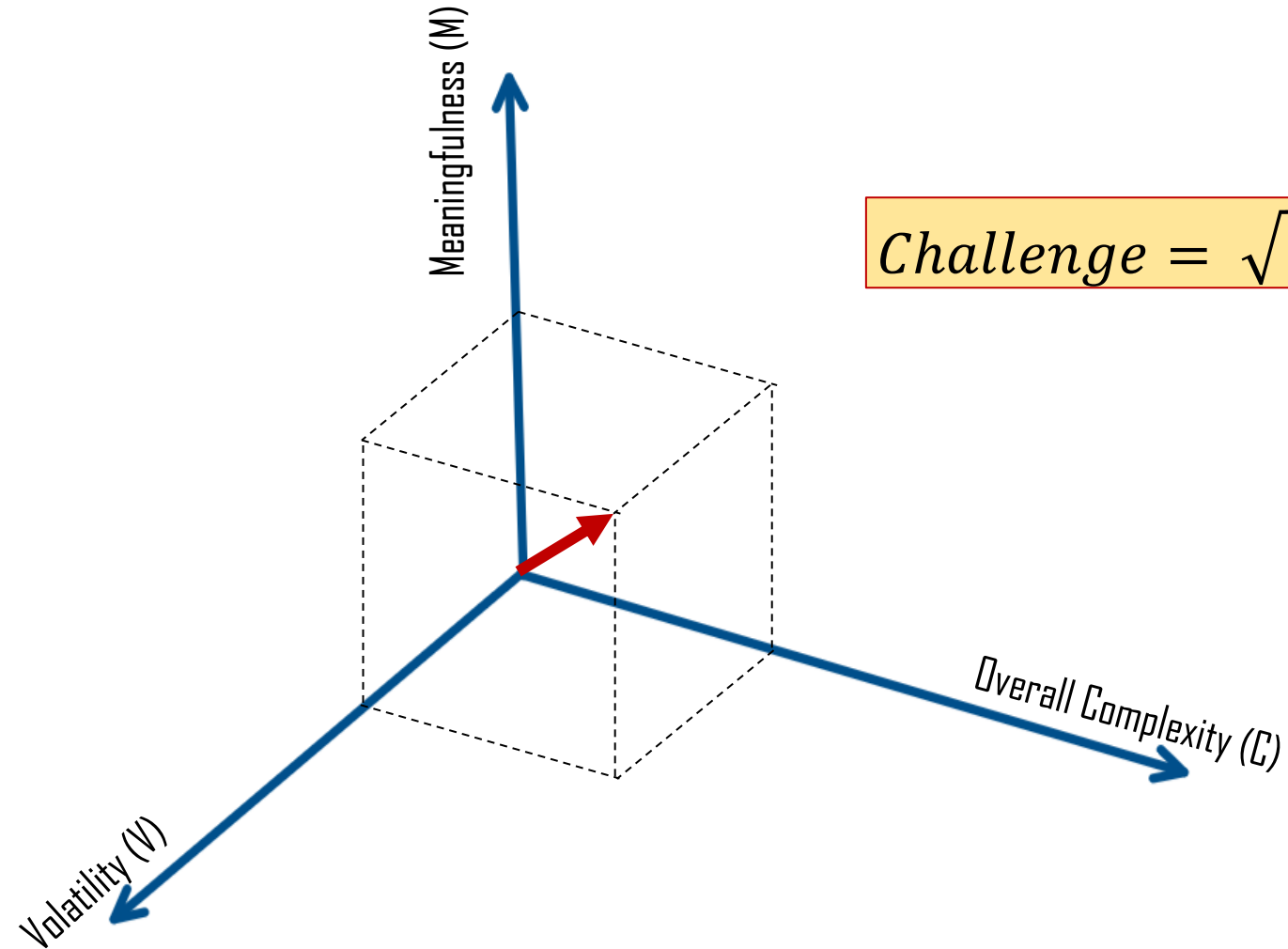
Unsupervised Learning
Semi-supervised Learning
Supervised Learning
Reinforcement Learning

Acceptable
Feasible
Suitable
Sustainable
Adaptable

References:

- Torres, M.T., Hart, G., Emery, T., 2019, The DSTL Biscuit Book: Artificial Intelligence, Data Science and (mostly) Machine Learning, 1st Edition revised v1_2
- The Royal College of Defence Studies, Getting Strategy Right (Enough), 2016.
- Smith, R., (2007). The utility of force. The art of war in the modern world. Alfred A. Knopf.

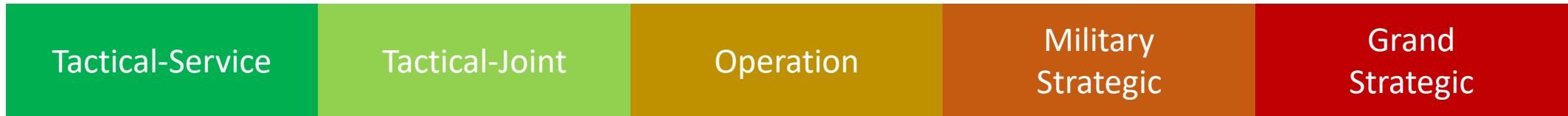
Challenge



$$Challenge = \sqrt{C^2 + M^2 + V^2}$$



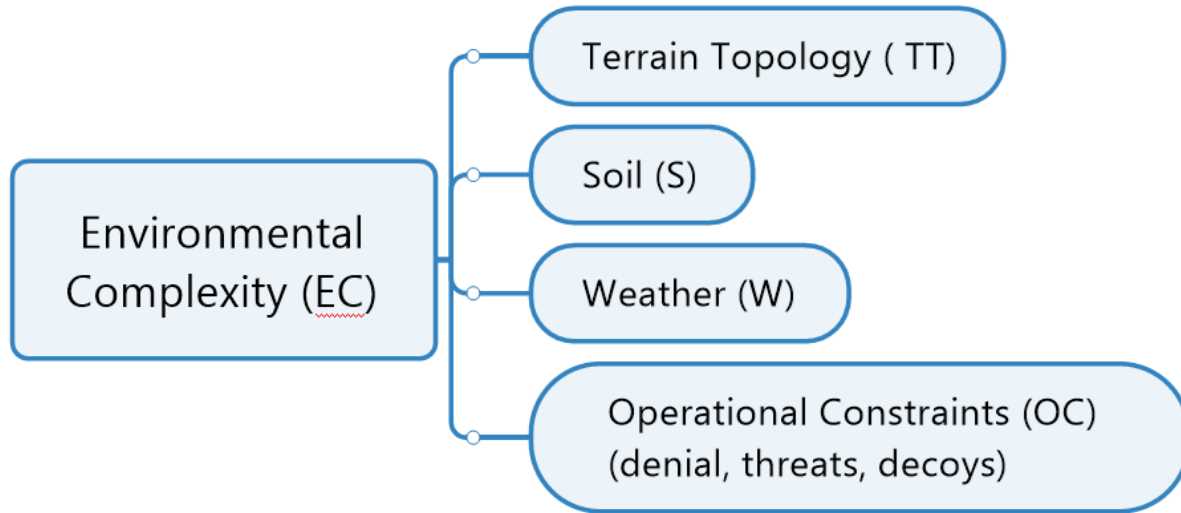
Meaningfulness (M)



References:

- o Joint Concept Note (JCN) 2/17, Future of Command and Control

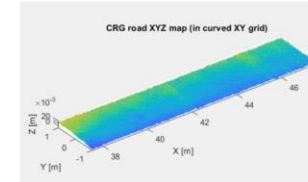
Environmental Complexity



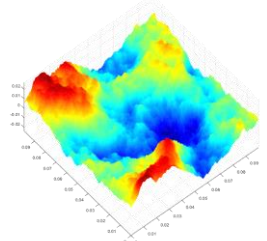
$$EC = f(TT, S, W, OC)$$

References:

- Huang, H.M., Pavak, K., Albus, J. and Messina, E., 2005, May. Autonomy levels for unmanned systems (alfus) framework: An update. In *Unmanned Ground Vehicle Technology VII* (Vol. 5804, pp. 439-448). International Society for Optics and Photonics.
- Wong, J.Y., 2008. *Theory of ground vehicles*. John Wiley & Sons.



$$S_g(\psi) = C_{sp}\psi^{-N}$$



Paved Roads

Rough Terrain

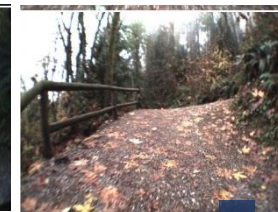
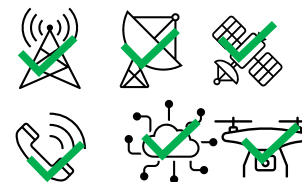
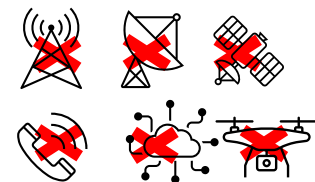


$$RCI = 2.6265 \frac{W(1-\frac{\delta}{h})^{1.5}}{bd^{0.4}} \cdot \left(\frac{s}{z^3}\right)^{0.2}$$



Hard Soil

Soft Soil



Mission Complexity

Mission Complexity (MC)

Command Structure (CS)

Knowledge Requirement (KR)

Collaboration Requirement (CR)

Dynamic Planning (DP)

Situational Awareness (SA)

Conflicted C2

Coordinated C2

Edge C2

Low

High

Stand-alone

Very High Collaboration

Routine Task

Dynamic Task

Very High SA

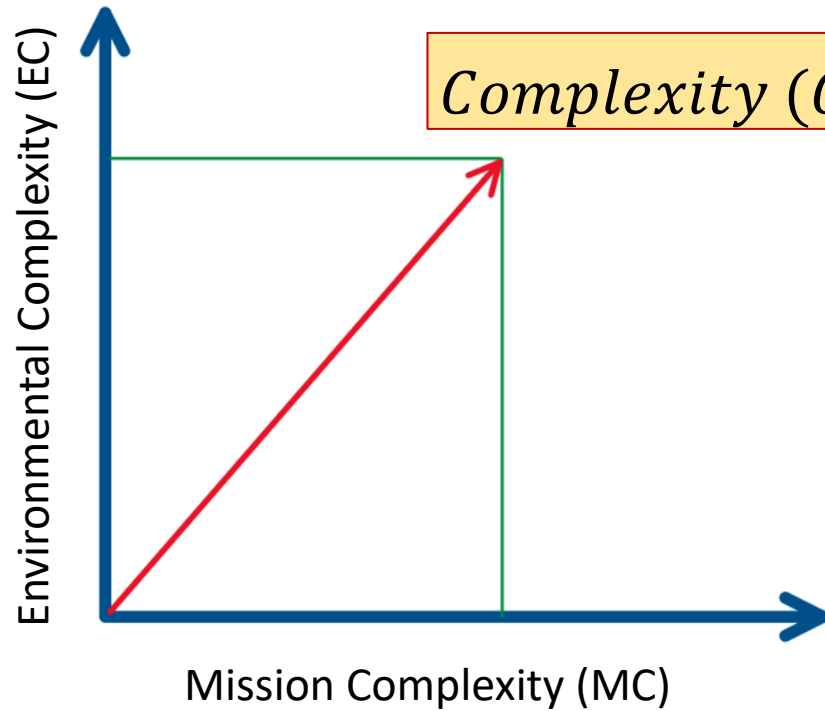
Operating Blind

$$MC = f(CS, KR, CR, DP, SA)$$

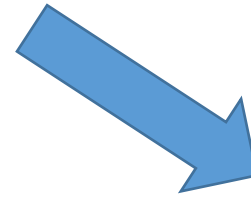
References:

- Huang, H.M., Pavsek, K., Albus, J. and Messina, E., 2005, May. Autonomy levels for unmanned systems (alfus) framework: An update. In *Unmanned Ground Vehicle Technology VII* (Vol. 5804, pp. 439-448). International Society for Optics and Photonics.

Complexity and Volatility



$$Complexity (C) = \sqrt{MC^2 + EC^2}$$



$$Volatility (V) = \frac{dC}{dt}$$

Section #1

Section #2

Section #3

Trustworthiness and AI

(An Account of Trustworthiness fit for Purpose)

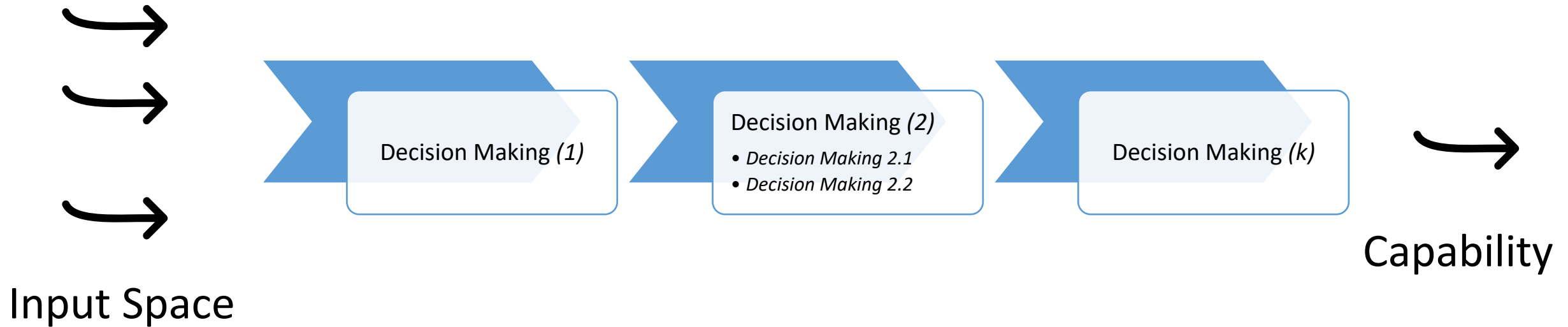
Trustworthiness Framework

(Technical Specification)

Enactment of Framework

(Fuzzy Logic)

Hierarchical Decision Making Topology

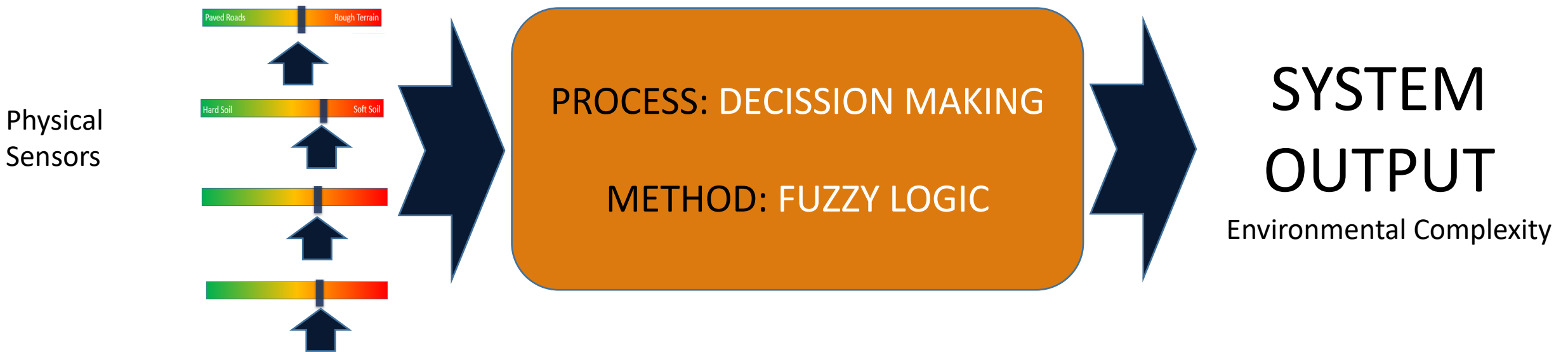


$$Capability = K \times (Means)^p \times (Ways)^q \times (Suitability)^r$$

Linguistic Expressions

Example – Environmental Complexity (data flow)

SYSTEM INPUT RANGES



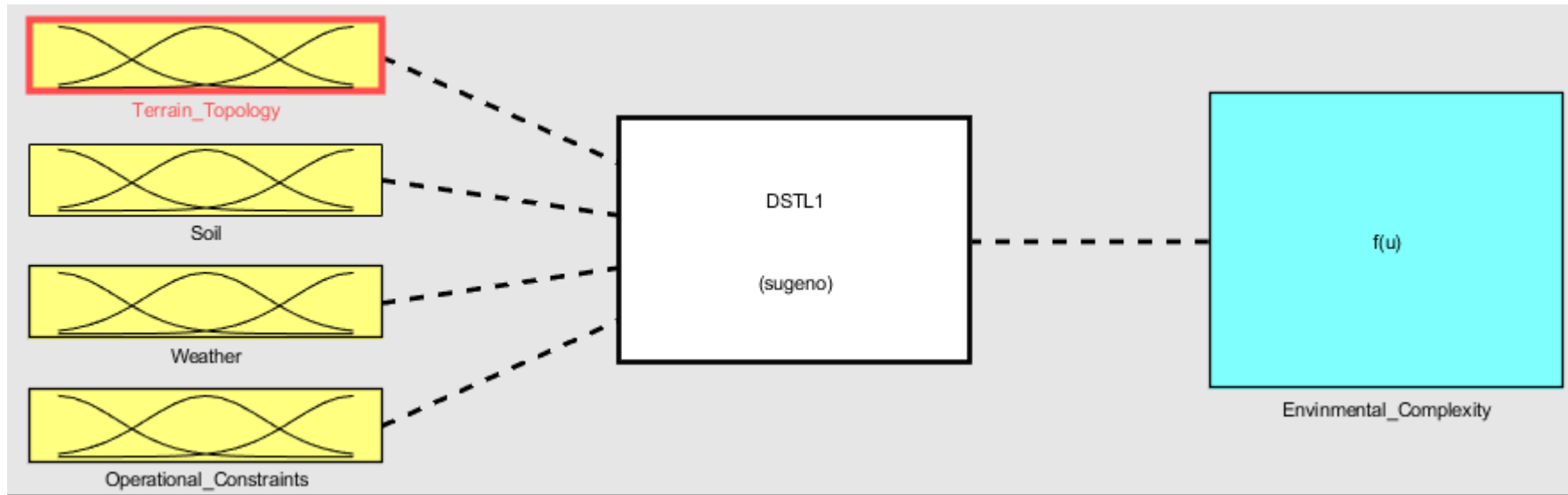
Environmental Complexity – Methodology (Block diagram)

Defining:

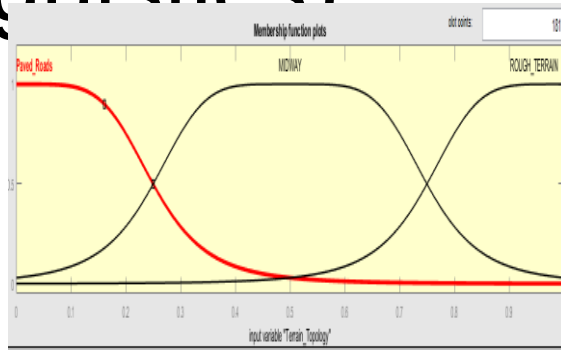
The number of input variables, 4

The input range for each variable, [0,1]

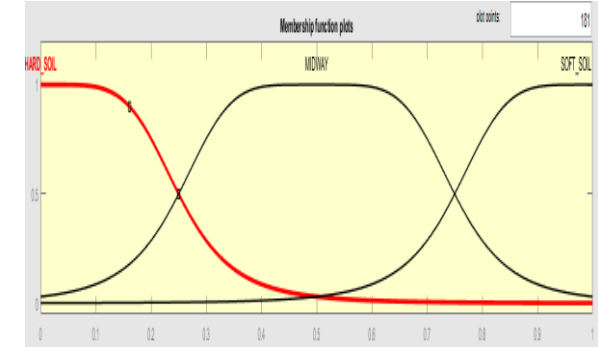
The number of subranges per range (membership functions), 3
resulting in 81 rules !



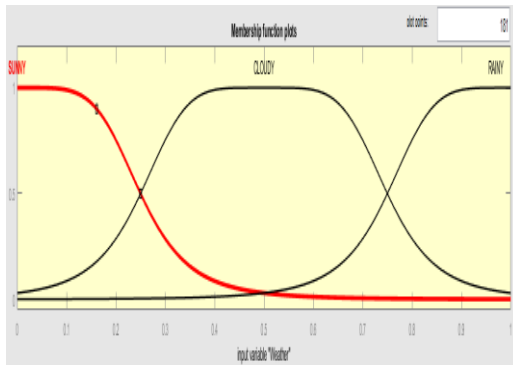
Example – Environmental Complexity (numbers, linguistics)



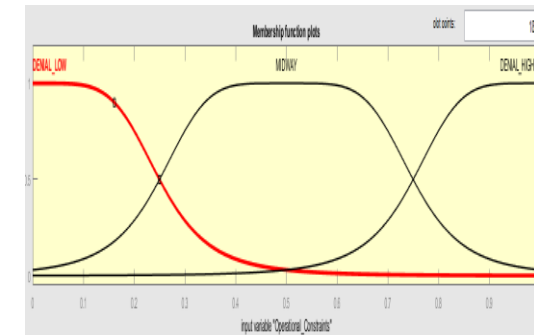
Terrain Topology (TT)



Soil (S)



Weather (W)

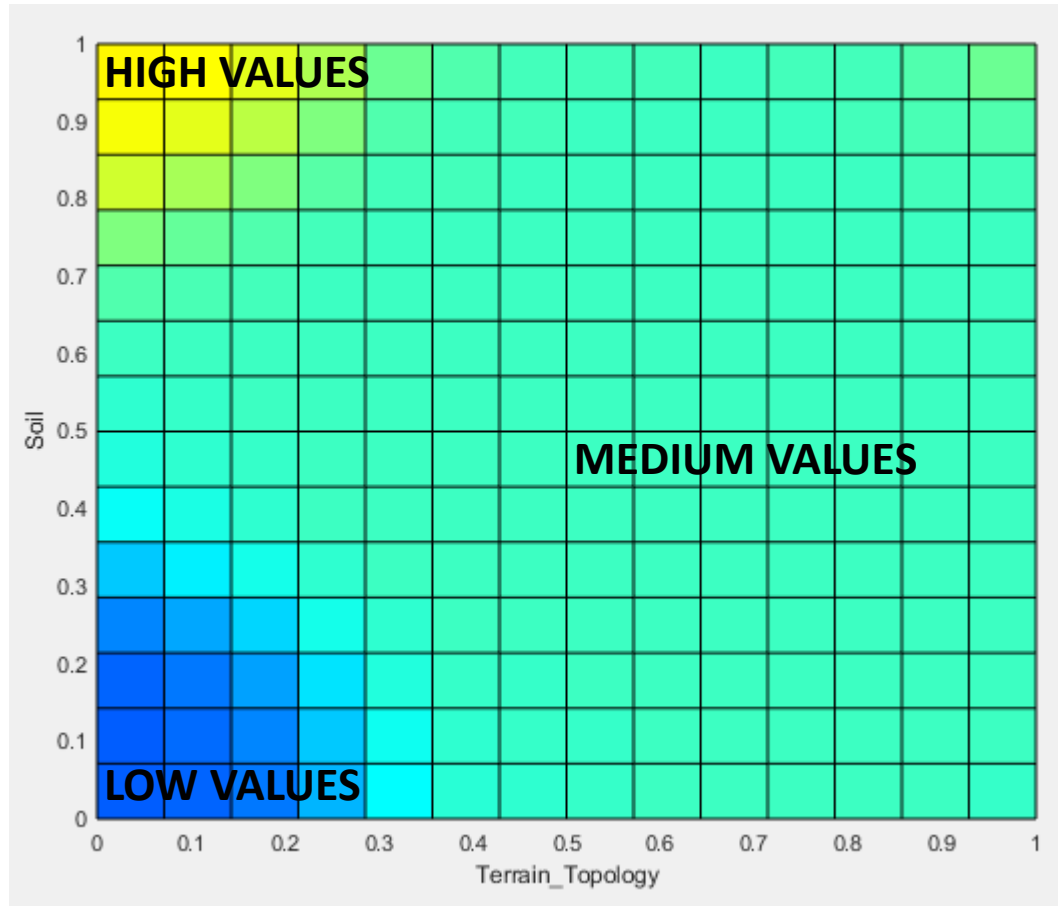


Operational Constraints (OC)
(denial, threats, decoys)

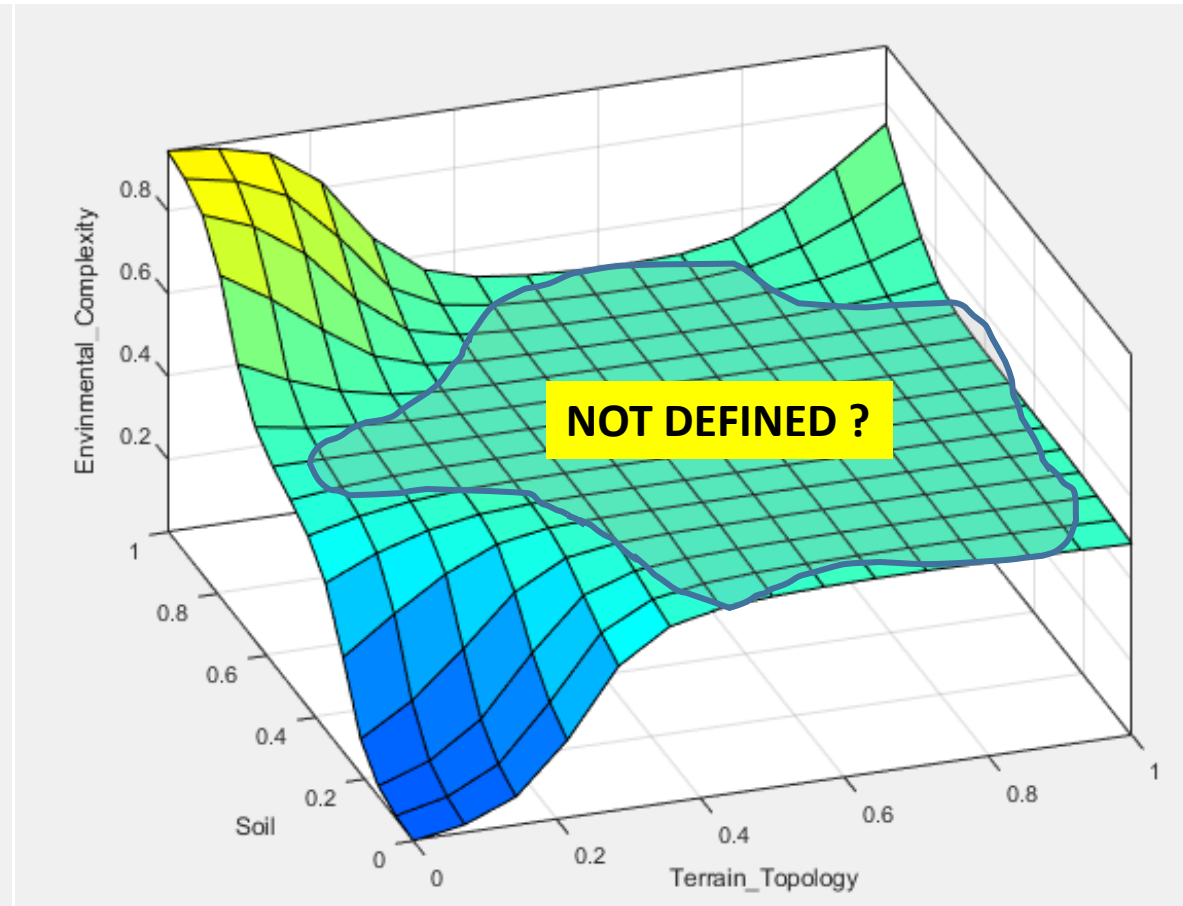


Environmental Complexity – Output Space

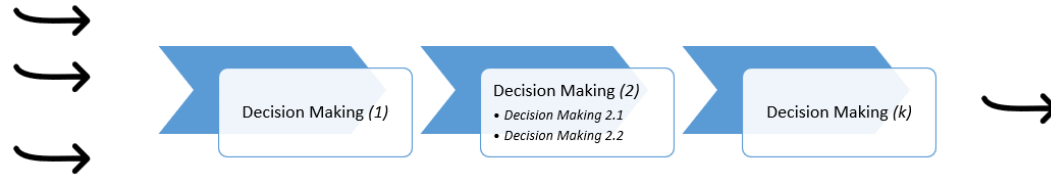
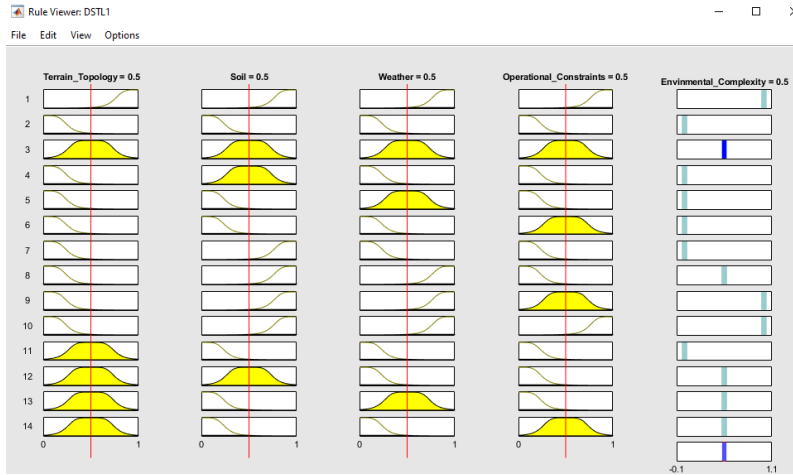
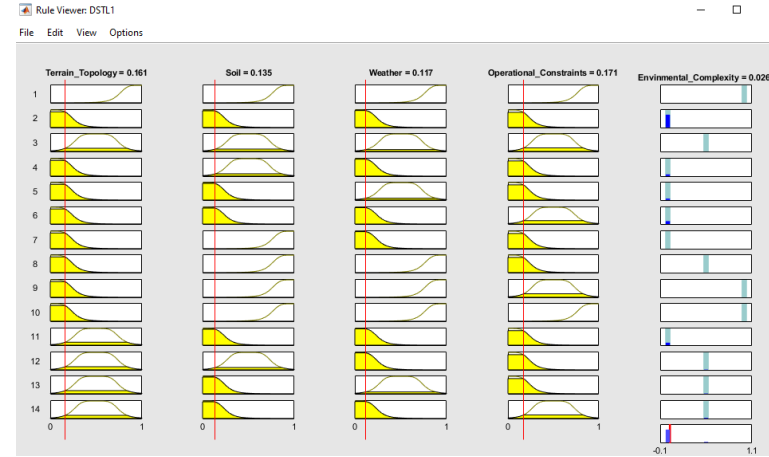
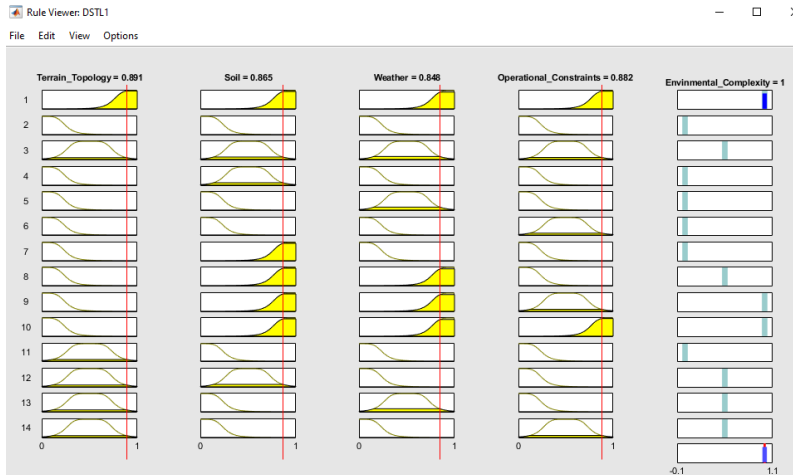
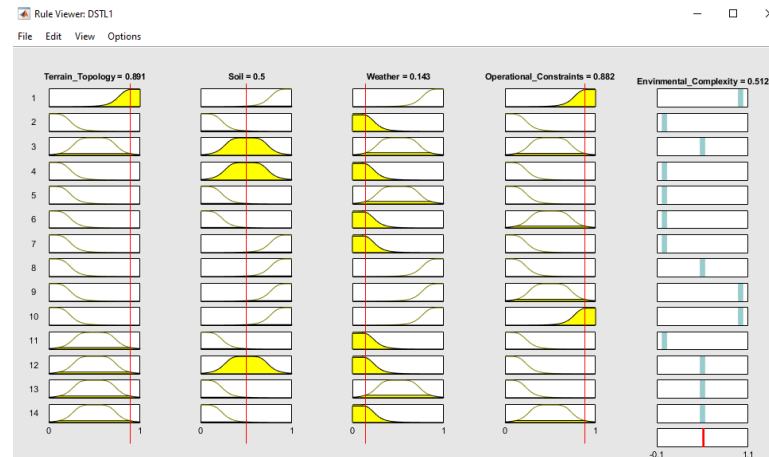
2D approach



3D approach



Hierarchical Decision Making - Reactivity





Conclusions

- The framework renders a transparent relationship between the AI's capability and suitability to the challenge it is tasked. This approach paves the way for debate on ethical and legal concerns.
- The framework provides a structure that can be directly applied across all five domains of defence (space, air, land, sea and cyber).
- The framework provides an amalgamation of objective measures and subjective judgments.
- The framework can be simply understood and interpreted.
- The variables of the framework are combined using technical (transparent and traceable) means to capture the complexities of the nuances between the variables into a coherent picture of trustworthiness.

Supporting the Dependency Relationship between Human-AI-Teaming promoting Trustworthiness.



DSPACE

<https://dspace.org/>

A novel framework for identifying the threshold of trust-worthiness of Artificial Intelligence

Forsyth, Tim; Kumar, Ajay; Economou, John T.

2021-10-12

Cranfield University

<https://www.gov.uk/government/news/global-ai-experts-to-present-at-dstls-virtual-ai-fest>

Forsyth T, Kumar A, Economou J. (2021) A novel framework for identifying the threshold of trust-worthiness of Artificial Intelligence. Presented at: AI Fest 4: Trustworthy AI, 12-13 October 2021, Virtual Event