

# Decision Maker Evaluation Framework

Gareth Rees and Claire Milner  
MBDA, Golf Course Lane, Filton, BS34 7QW

## Abstract

*This paper proposes a Decision Making Evaluation Framework (DMEF), to show potential for an approach that allows a quantitative evaluation and comparison of decision technologies and the relationship between decision performance and characteristics of context of use. Building on previous SEAS DTC work (particularly OA005) the DMEF is applied to classifier / inference and route and path planning decision problems. Using established hypothesis and criteria, evaluations were conducted, to show that: the DMEF supports algorithm tuning; allows assessment of a technology's robustness; enables system-trade-offs between decision making, sensing and control; and informs decision chain architectural choices.*

Keywords: Evaluation, Framework, Autonomous Systems, Systems Engineering, Decision Maker

## 1. Introduction

Historically there has been a great deal of work investigating automated decision making from a wide variety of diverse fields, e.g. probability theory, statistics, control theory, economics, pattern recognition, AI, and game-theory. This has produced a wide variety of relevant decision making technologies including: planners, statistical classifiers, Dempster Shaffer theory, Bayesian Belief Networks (BBNs), multi-criteria decision analysis, Markov Decision Processes (MDPs), Partially-Observed MDPs (POMDPs), and Rational Agents [1].

However, while there is tacit understanding of the assumption, strengths and weaknesses for many of these decision technologies, these are often focused on optimisation within a particular problem instance.

From a systems designer's perspective, which technologies fit to a particular application domain is not well understood. In particular, as real systems operate within a wide range of conditions, optimality

becomes an academic issue and 'good enough' solutions with understood limitations that can be applied across a range of problems are to be preferred.

The performance of a decision maker is also critically dependent on the character of the system context within which it resides. Consequently, the system design problem spans not only algorithm technology selection but also the performance of the supporting sensor, actuation and communication subsystems. Furthermore, often the output from one decision maker informs subsequent decisions, so the design must also address the architecture used to combine information or decision flows.

While many technologies recognise these dependencies, there has been comparatively little previous work to characterise quantitatively the performance of decision technologies in relation to the sensor and actuation uncertainties and the decision architecture.

The challenge addressed in the SER018 project, therefore, is to be able to exploit the considerable body of decision

technologies by providing the quantitative characterisation that is key to informing design decisions.

In particular this paper contributes to addressing this challenge by describing the evaluation framework needed to characterise decision makers. In this context an evaluation framework is used to mean: an approach that allows quantitative evaluation and comparison of decision technologies and the relationship between decision performance and characteristics of the context of use.

In this paper we describe the proposed Decision Making Evaluation Framework (DMEF) and show, by application to illustrative ‘toy’ problems using a ‘grab-bag’ of readily available technologies, how the DMEF can be used to evaluate technologies and support the wider system design activities.

It should be noted that the SER018 project started in January 2009 and, at the time of writing (March 2009), is still a work in progress. This paper is intended to provide a snapshot of initial work in order to highlight the goals and approach.

## 2. Overview of the Paper

The remainder of this paper is constructed as follows:

- **Section 3** highlights earlier work undertaken within the SEAS-DTC OA005 project that provides a Decision Making Framework (DMF) for identifying candidate decisions for allocation to autonomous systems. Some of these identified decisions are used as prototypes for model problems for the evaluation of decision technologies in the remainder of this paper.
- **Section 4** proposes a DMEF, based on the CMU evaluation process [2]. This supports the quantitative evaluation of candidate decision making algorithms

and technologies such that the relative benefits of decision making technologies versus other design choices (such as better sensing or better actuation performance) can be understood.

- **Section 5** sets out initial criteria for evaluation and visualisation of decision maker performance. This focuses on evaluating the correctness of the decision in terms of performance and risk.
- **Section 6** provides a case study of evaluation applied to detection, recognition and identification problems highlighted in the DMF.
- **Section 7** provides a case study of evaluation applied to route and path planning problems highlighted in the DMF.
- **Section 8** presents a summary, conclusions and route forward.

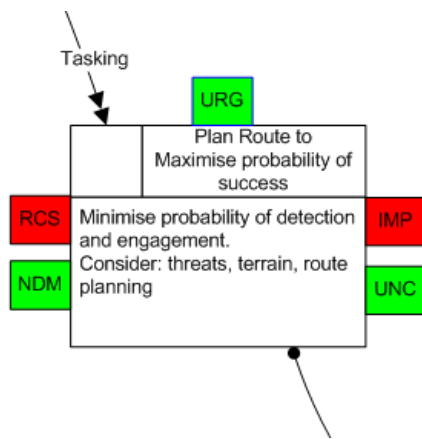
## 3. Background – DMF

The work reported here seeks to build on the body of previous work within the SEAS DTC, e.g. SER012, MP001, AA012, AA016, OA005 and MAS – RP07321. In particular, the Decision Making Framework generated by CORDA [4] is used as a launch point for this work. This framework ‘*examines the range and types of decision, and the relationships between them, in the planning and management of operational missions*’. This framework was applied to the SEAS hostage situation and Suppression of Enemy Air Defences (SEAD) vignettes to identify and describe decisions, the connecting flows and provide a decision classification in terms of five attributes:

- **Urgency:** This gives an indication of the time available to make each decision.
- **RCS Complexity:** This indicates the complexity of the decision in terms of Rational Choice Strategy (RCS).

- **NDM Complexity:** This indicates the complexity of the decision in terms of Naturalist Decision Making (NDM).
- **Importance:** This indicates the impact of the decision on the mission objectives and constraints.
- **Uncertainty:** This indicates the uncertainty of the information available about the situation.

The following figure, taken from the CORDA analysis of the SEAD vignette, illustrates the classification of an example decision against each of these attributes in terms of: high (red), medium (amber) and low (green) traffic lights.



**Figure 1: DMF decision classification example from the SEAD vignette analysis**

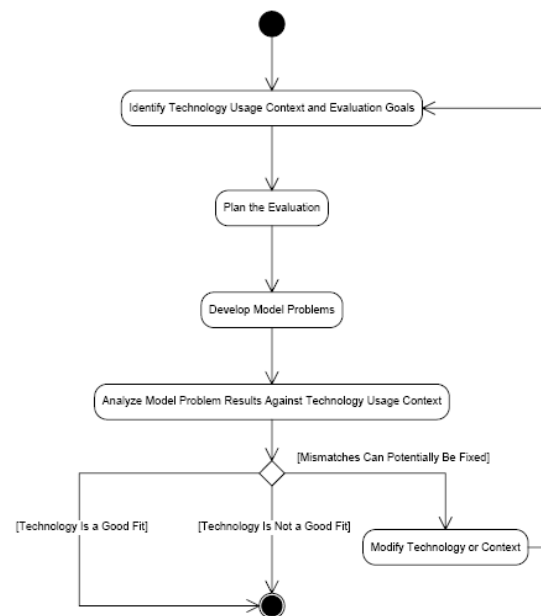
This classification was then used as basis for allocating decisions between humans and autonomous systems on the basis of six hypotheses, e.g. ‘1. Complex RCS decisions are more effectively carried out by autonomous systems’. Thus, this framework offers a route to analyse problem vignettes to generate and classify decision architectures in order to identify those candidate decisions where the application of autonomy would be most appropriate and beneficial.

However, the DMF uses a somewhat subjective and coarse granularity. Also it does not address the mapping between decisions and technologies. The DMEF defined in the current paper compliments and builds on the DMF work to develop a

quantitative framework that can be applied for comparison of technologies in order to guide the system design activities.

#### 4. The Decision Making Evaluation Framework (DMEF)

The evaluation framework proposed here builds on the CMU Context-Based Technology Evaluation Process [2]. This describes a process that determines the fitness of a technology by experimentation within a specific context. The process is illustrated in Figure 2. This is expanded in relation to decision maker evaluation in the following subsections.



**Figure 2: CMU Context-based technology evaluation process**

##### 4.1 Usage Context and Evaluation Goals

The goal of this activity is to determine the context and goals for conducting an evaluation.

As highlighted in the introduction, the context here is the design of a future autonomous system. The DMF framework provided many possible example contexts for evaluation. However it was felt that these were too specific / detailed and that the authors’ of the current work didn’t have access to the complete context for these

decisions. Rather, it was felt that more general or prototypical decision problems spanning aspects of the whole problem space were needed.

In order to identify such prototypical problems an initial clustering activity was undertaken using the results from OA005. Due to limited detail this should be regarded as indicative rather than definitive. This initial partition focused on finding ‘clusters’ of decision making problem types:

- Various detection, recognition identification, situation assessment, and threat evaluation problems were identified. These were recognised as all being examples of object classification or inference problems [5].
- Various decisions to plan or re-plan sequences of states and actions were identified. These were recognised as examples of Sequential Decision Making (SDM) or planning problems [6]

**Table 1: Decision Problem Clusters**

Problem Type	Examples
Classification / inference	2.2 Determine the nature of sites from imagery 4a.3 Have we detected a SAM site 4a.6 Have we classified a threat
SDM, Planning	1.1 Plan route to maximise probability of success 4a.1 Decide how to navigate to potential SAM site

Other types of decision problem, such as dynamic reasoning, constraint satisfaction, scheduling / assignment and association problems were also identified. However, these decision types occurred in smaller numbers.

Examples of these decision types are given in Table 1. Notice that in reality some of the highlighted decisions may contain elements of several problem classes.

It was decided that initial work should assume these two problem classes as contexts for evaluation. This focus is reflected in the content of sections 6 and 7.

Within these contexts there may be many reasons for undertaking a technology evaluation. It is important that the evaluation be aware of these reasons, so as to ask and answer the right questions. The following examples, cast around the OA005 attributes, illustrate how these might set the goals for decision maker evaluation:

- *Urgency*: The issue is execution time for various technology options and the speed of human interaction.
- *RCS Complexity*: The issue is the completeness / exhaustiveness of the approach and the impact of curtailing the search before convergence.
- *NDM Complexity*: The issue is the diversity of required encyclopaedic information and the approach for its application to the individual decisions.
- *Importance*: The issue is correctness in terms of the balance between system performance and operational risk.
- *Uncertainty*: The issue is the robustness of candidate technologies to errors in assumptions about the operational environment.

In addition, there are many questions regarding the architecture of the decision maker [7]. The issue here is the trade-off between the structure of the decision maker and its performance.

In order to provide a simple but representative setting, it was decided to focus on the evaluation of:

- The performance of different technologies within the highlighted contexts.

- The sensitivities and robustness to uncertainty (including errors in the assumed level of uncertainty).
- The trade-off between better decision making and further information from better / additional sensors or from better actuation.
- The impact of decision maker architecture on decision performance.

#### 4.2 Evaluation Planning

The goal of this activity is to plan the practical aspects of evaluation. In addition to the obvious management factors (e.g. team, stakeholders, effort, schedule) this planning activity should consider the provision of baseline / benchmark technologies and datasets needed for evaluation.

#### 4.3 Model Problem Development

An integral part of the process is the development of model problems. The steps for developing model problems are as follows:

1. **Develop hypotheses.** Hypotheses are claims regarding the fitness of the technology / architecture that may be sustained or refuted.
2. **Develop criteria.** Criteria are used to determine if a model solution sustains or refutes a hypothesis. Criteria are measurable statements of capability. Each hypothesis can be associated with one or more criteria.
3. **Design model solution.** A model solution is the simplest set of components that are able to answer the questions posed by the hypotheses and associated criteria, within the required context.
4. **Implement and evaluate model solution against criteria.** In this activity, the model solution is implemented, run, and observed, until there is enough information to sustain

or refute the set of hypotheses.

The criteria used within the evaluation are critical to the above process. Initial criteria appropriate for the evaluation of decision making technology are given in Section 5.

#### 4.4 Results Analysis

The last step in the process is to make a judgement with respect to the fitness of the technology for the context of use. The results from the model problems provide informed design choices because of the direct evidence from experiments. Based on the results of the evaluation, and the outcomes of testing the chosen hypotheses, there should be evidence to determine the degree of fit:

- If the model problem provides evidence that supports the set of hypotheses, assuming sufficient confidence in the results, the technology can be declared a good fit.
- If the model problem provides evidence that refutes the set of hypotheses, it can be stated that the technology is not a good fit.
- If the model problem provides evidence of performance deficiencies that might be solved by modifying the technology or context, then the evaluation should be repeated with such changes.

### 5. Criteria for Visualising and Evaluating Decision Making

The natural framework for evaluation of decisions is decision theory [1]. This seeks to maximise or minimise (depending on context) Expected Utility, EU, of outcomes  $O_i$  from the decision for action / policy  $A$ :

$$EU(A) = \sum_i P(O_i(A)) \cdot U(O_i(A))$$

where  $P(O_i(A))$  are the probabilities of the individual outcomes and  $U(O_i(A))$  is a function that maps outcomes to utilities /

costs that encodes preferences among outcomes.

Most often  $O_i(A) = [o_1(A), o_2(A) \dots o_n(A)]$  is a vector where the outcome has many attributes. Multi-attribute decision theory [1] considers utility as the weighted sum or product of individual attributes.

EU is often used within decision technologies as a basis for a multi dimensional search which seeks to optimise the EU over decision options / parameters. From the designers perspective this entangles parameters and performance in two unhelpful ways: i) optimality is not necessarily required, ii) even an optimal answer may not be good enough for use.

Furthermore, methods often seek to optimise a cost function that combines utilities as free parameters to capture generally desirable and undesirable outcomes, or to configure additional intermediate attributes relevant to how the problem is solved.

This means that the utility reported by a method does not generally relate to utility in the real world, where utility is known to be difficult to capture [8].

The assumption of linear and mutual preferential independence (MPI) utility allows performance attributes and utility to be separated such that:

$$EU(A) = \mathbf{U} \cdot \bar{O}(A)$$

Where the utility function is now a vector of weights  $\mathbf{U} = [u_1, u_2 \dots u_n]$  and  $\bar{O}(A) = [\bar{o}_1(A), \bar{o}_2(A), \dots \bar{o}_n(A)]'$  is the expectation over outcomes from action / policy  $A$  calculated from the experimental results. The MPI simplification is usually appropriate for the system design stage where the utilities are most often preferentially independent.

By separating performance from utility [9] in this way, the decision goal can be projected into a Cartesian space for visualisation and analysis where two points have the same expected utility if:

$$\frac{\bar{o}_1(A) - \bar{o}_1(B)}{\bar{o}_2(A) - \bar{o}_2(B)} = \frac{u_2}{u_1}$$

This defines the gradient of lines of iso-EU which can be used to select, compare and trade-off decision performance in terms of multiple attributes.

In general, the success of a system design will be determined by the extent to which it is possible to maximise the intended performance attributes while minimising the risk of undesired attributes.

This suggests that the most informative visualisation and evaluation will be found by projecting into a space that represents the tension between the desired performance and the undesired attributes as illustrated in Figure 3.

Figure 3 shows a blue backdrop describing the utility. The reward/utility structure would be determined by the decision context (see Sections 6 and 7). Also displayed are Diamonds representing key points: Blue: Zero performance – zero risk, Green: 100% performance – zero risk, Red: 100% performance – 100% risk.

Given a choice between the points  $a, b, c$ , point  $b$  is clearly preferred as it dominates  $a$  in terms of performance and  $c$  in terms of risk. Consequently, the performance characteristics of a decision can easily be understood with this performance-risk space either as points or as a curve.

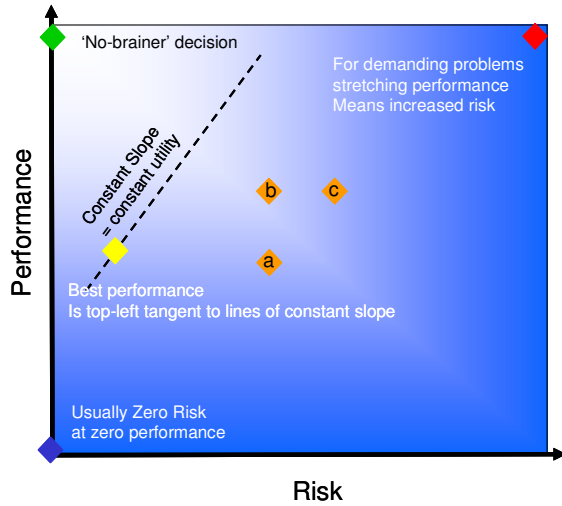


Figure 3: Illustration of the decision space

The use of this performance-risk space is applied to the highlighted contexts in the following subsections.

### 5.1 Classification / Inference

For classification, the principal desired performance criteria is the probability of correct classification / inference regarding the object state,  $\bar{o}_1 = P_{TP}$ , while the principal undesired performance is the probability that objects are misclassified,  $\bar{o}_2 = P_{FP}$ . The use of ROC graphs to visualise and evaluate classifiers in this way is well understood [3, 9, 10] and is adopted as the criteria for the evaluation of classification in the next section.

### 5.2 SMD / Planning

For path planning, it is proposed that the principal desired performance attribute is optimisation of properties of the selected path or route,  $\bar{o}_1 = \lambda$ , e.g. the path length, while the undesired performance attribute is the probability for some undesired event,  $\bar{o}_2 = P_F$ , e.g. the vehicle crashes. These attributes are calculated from the outcomes of  $n$  experiments within an evaluation as follows:

$$\lambda = \frac{\sum_{i=1}^n \lambda_i(A)}{n}, \quad P_F = \frac{\sum_{i=1}^n H_i(A)}{n}$$

Where  $\lambda_i(A)$  is the desired property of the path obtained from the  $i^{th}$  outcome from application of policy A and  $H_i(A)$  is an indicator function that indicates a failure event within the  $i^{th}$  outcome.

This new Planning Operating Characteristic (POC) is adopted as the criteria for the visualisation and evaluation of path planning and routing in section 7.

## 6. Classifier / Inference Decisions

The evaluation of classification / inference decisions focused on testing the following hypotheses:

- It is possible to tune and select different classification technologies.
- It is possible to evaluate the sensitivities and robustness of the decision maker to sensor uncertainty.
- It is possible to evaluate the value of additional information from a better sensor.
- It is possible to compare decision maker architectures.

Consistent with the desire to maintain simple model problems, the evaluation used synthetic data comprised of Gaussian mixtures that were operated on by Linear Discriminant and Quadratic classifiers [5]. The free parameters were the true and assumed class conditional sensor distributions, class priors and information architecture. In each case ROC graphs were produced to visualise and analyse the hypothesis.

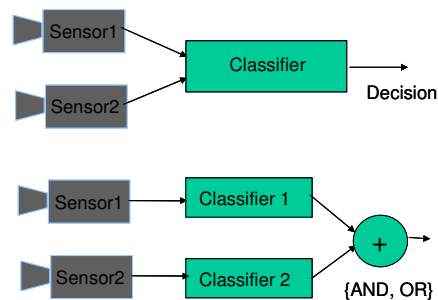
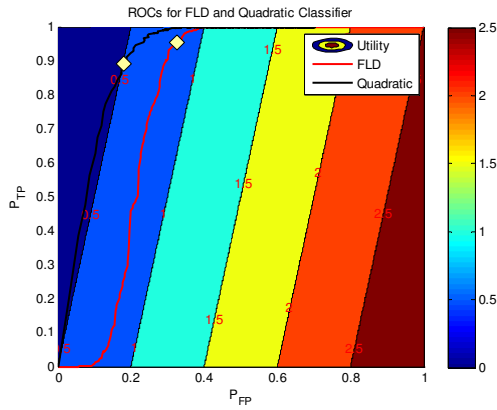


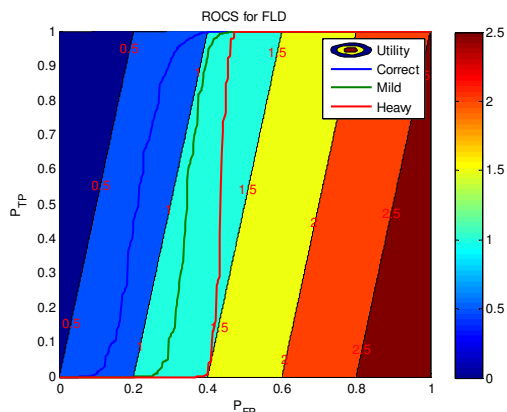
Figure 4: Data fusion and hierarchical decision architectures



**Figure 5: Algorithm selection and tuning**

Figure 5 shows results for two classifier technologies applied to a two class problem with 1:5 ratios on the utility of missed and false positives (the resulting utility is displayed on the figure background). It is clear that the graphs support the hypothesis as the parameters for the technologies can be tuned to maximise utility (yellow diamonds) and the quadratic classifier would clearly offer better performance.

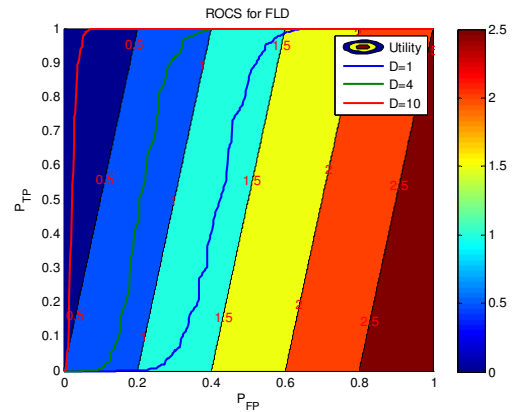
Figure 6 shows results for a classifier where the algorithm assumptions have errors with respect to the true sensor response model. Here the true sensor model for clutter has varying degree of ‘tail heavy’ statistics. This result supports the hypothesis that robustness can be evaluated as it clearly shows that the classifier is not robust to such uncertainty.



**Figure 6: Robustness to errors in sensor parameters**

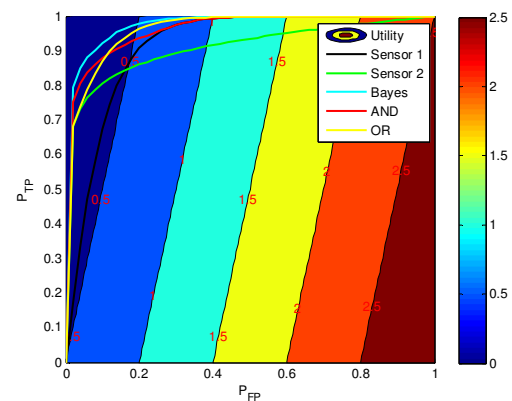
Figure 7 shows results for varying sensor performance, where  $D$  is the class

separation afforded by the sensor. It is possible to compare these results with the results for changing classifier technology supporting the hypothesis that it is possible to evaluate trade-offs between subsystems and decision technologies.



**Figure 7: Variation with sensor performance**

Figure 8 shows the performance of two individual sensors together with three decision architectures: (i) hierarchical AND of local decisions, (i) hierarchical OR of local decisions, and the results of a single Bayesian fusion. These architectures are illustrated in Figure 4. It is possible to compare the results supporting the hypothesis that it is possible to evaluate trade-offs between architectures.



**Figure 8: Impact of decision architecture**

In all cases the evaluations have supported the hypotheses. In a real application it would be possible to evaluate the fitness of combinations of technologies, subsystems and architectures.

## 7. Route and Path Planning

An MDP [1] and a road-map planner [6] were applied to the ‘toy’ route planning problem shown in Figure 9.

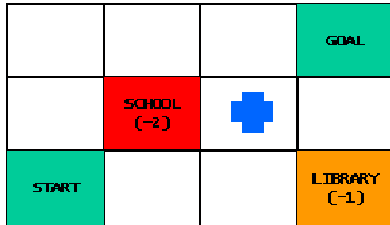


Figure 9: The model world used for route-planning investigations

This contains uncertain control that can cause the vehicle to crash into the school or library and includes uncertain sensing of a possible ‘mine’ (denoted with a blue cross) with an associated ROC characteristic. The DMEF looked to investigate the following hypotheses:

- It is possible to evaluate various route planning technologies.
- It is possible to evaluate the system considerations such as control and uncertainty in the system.
- It is possible to evaluate new system concepts including system additions / improvements.
- It is possible to evaluate how the sensitivity to information impacts the decision and to what extent the uncertainty should be trusted.
- It is possible to evaluate the impact of classification performance and planning performance.

Figure 10 shows the results of four separate routes from the start to the goal. The utility is set to a 1:5 ratio favouring time restrictions than risk. Using the utility it is possible to evaluate the route planner technologies hence verifying the first hypothesis.

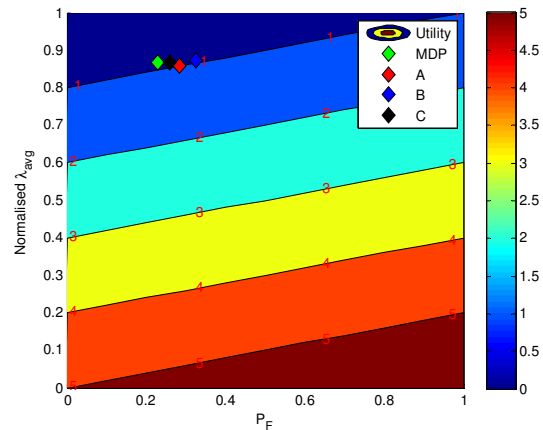


Figure 10: Technology analysis

Figure 11 shows the results when varying the control uncertainty parameter of a single route planner. It is possible to compare the results according to a utility such as shown in figure 10 hence supporting the hypotheses to evaluate system considerations.

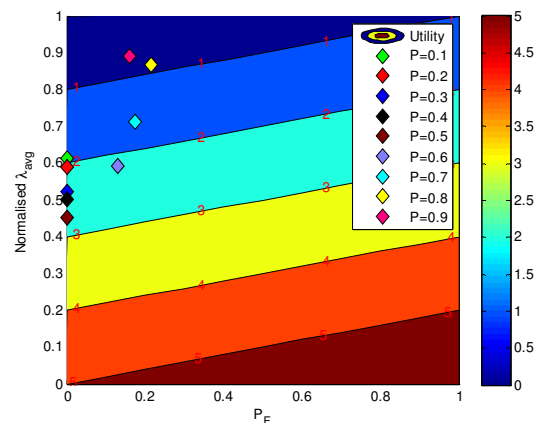


Figure 11: Control Analysis

Figure 12 shows the results when looking at system design choices using a single route planner. System tradeoffs such as higher speed lower control (i.e. adding a booster to the system) are considered and evaluated. This confirms the hypothesis that it is possible to evaluate system design choices.

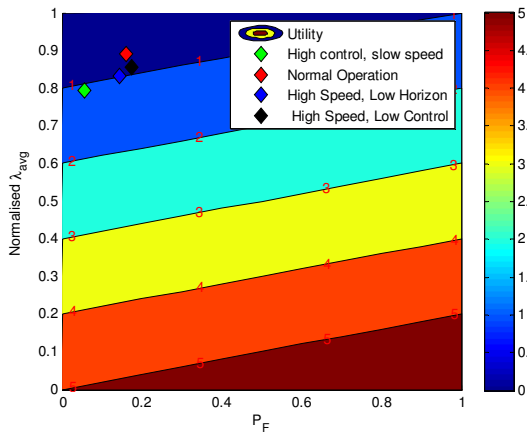


Figure 12: System design analysis

Figure 13 shows the results as the assumed parameters varies from the actual environmental conditions. This shows that it is possible to evaluate the best route according to each condition hence supporting the test hypothesis.

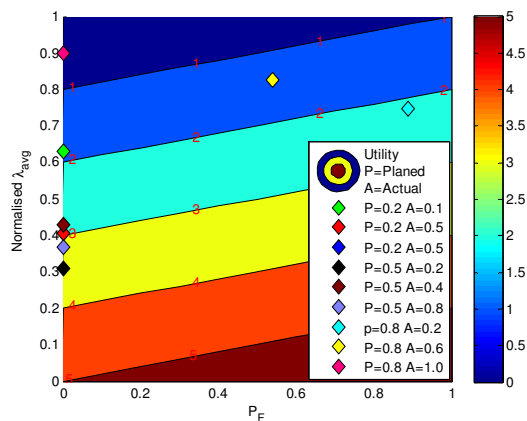


Figure 13: Sensitivity analysis

Figure 14 shows the results as the performance of the sensor detecting the mine is varied according to the ROC characteristic for the linear discriminant classifier from Figure 5. This shows it is possible to evaluate the impact of sensing performance on planning hence supporting the final hypothesis.

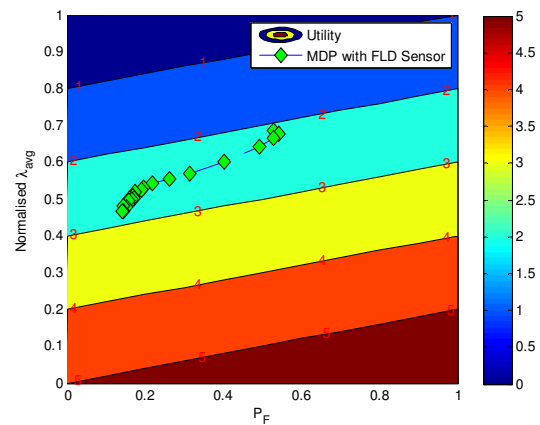


Figure 14: Classification impact on planning

In all cases, each of the evaluations has supported the hypothesis under test. Again, it would be possible within a real application to evaluate the fitness of combinations of technologies, platforms and system parameters.

## 8. Summary and Conclusions

This paper has proposed and described an evaluation framework for decision maker technologies. Building on previous DMF work, this provides an approach that allows quantitative evaluation and comparison of decision technologies and the relationship between decision performance and characteristics of the context of use. A key contribution here is the use of the existing ROC and its extension to new POC criteria that have enabled the visualisation and evaluation of performance

This has been an initial exploratory study to show that evaluation might provide useful inputs to inform the system design task. The framework has been applied illustrative ‘toy’ problems using a few readily available technologies and has shown that the DMEF:

- Supports algorithm tuning.
- Allows assessment of technology robustness.
- Enables system-trade-offs between decision making, sensing and control.

- Informs decision architectural choices.

This initial success suggests that the work progresses to problems of more realistic scale and complexity with a focus to support the TIED integration on the route to application within military autonomous system development programmes.

### References

- [1] Russell S. & Norvig P., *Artificial Intelligence: A Modern Approach*, 2<sup>nd</sup> Ed., 2003, Prentice Hall.
- [2] Lewis G.A. & Wrage L., *A Process for Context-Based Technology Evaluation*, 2005, CMU/SEI-2005-TN-025.
- [3] Swets J.A, Dawes R.M. & Monahan J, *Better Decisions through Science*, Oct 2000, Scientific American pp82-87.
- [4] Huges M., Tulip M. & Watson C. *Decision Making Framework*, May 2006, CORDA/CDR382/TR1.
- [5] Duda R.O., Hart P.E. Stork D.G. *Pattern classification*, 2<sup>nd</sup> Ed. 2001, John Wiley.
- [6] LaValle S., *Planning Algorithms*, 2006, Cambridge.
- [7] Deeks C., Vitanov I. & Williams W., *Safety Critical Autonomy*, 2008, SEAS-DTC Conference in Edinburgh.
- [8] Roth R & Field F., *Multi Attribute Utility Analysis*, 1994, J. Computer-Aided Materials Design, vol. 1, no. 3.
- [9] Fawcett T., Provost F., *Robust Classification for Imprecise Environments*, 2001, Machine learning Journal, vol42, no. 3.
- [10] Rees G.S., Wright W.A. & Greenway P., *ROC Method for the Evaluation of Multi-class Segmentation/Classification Algorithms with Infrared Imagery*, 2002, Proc. BMVC.

### Acknowledgements

The work reported in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence. The authors would also like to express their thanks to Mike Mew and Dr John Wadsworth (MBDA) and Dr Phil Greenway (SEIC).