

# Trusted Reasoning Engine for Autonomous Systems with an Interactive Demonstrator

We present a mathematical formalism for an explainable decision-making system and illustrate its applications in the context of a simple urban search scenario.

By  
**Jens Kröske, Kevin O'Holleran and Hannu Rajaniemi**  
 ThinkTank Maths Limited  
 www.thinktankmaths.com  
 Email: a.mathis@thinktankmaths.com

This paper was presented at the Electro Magnetic Remote Sensing (EMRS) Defence Technology Centre (DTC) conference, Edinburgh, July 2009. Please see www.emrsdtc.com for further information.

## Introduction

A key mission management issue in tasks involving autonomous systems is *human-machine trust*. Human operators can be required to supervise and collaborate with machines that often react to their environment in unexpected ways. To accept and trust autonomous systems, human operators need to be able to understand their reasoning process and the factors that precipitate certain actions. The machine needs to be able to communicate the reasoning behind its actions in an unambiguous manner that is also accessible to non-technical personnel – in other words, to *explain itself*.

While there has been some previous work on explainable artificial intelligence (AI), most of it has focused on static expert systems [1] or rule-based agent frameworks [2]. There has been relatively little work on explainable reasoning systems that operate in dynamic environments involving uncertainty.

The aim of the current project is to set a firm mathematical foundation for explainable decision-making systems and eventually produce an implementation of such a system, tailored to a particular scenario (e.g. urban search) – along with an interactive demonstrator allowing users to interact with autonomous systems that attempt to explain their actions.

At the time of writing, the project is approaching the end of its first six-month phase and has so far focused on exploring decision-making and explanation in the context of a simplified toy model.

## Explanation philosophy

Theory of explanations and explainability is a subject usually studied by philosophers and psychologists rather than

mathematicians or computer scientists. However, explainable decision-making requires us to capture some elements of an explanation formally – something that can be informed by a given philosophical approach to explanation [3]. It is therefore worth clarifying precisely what we mean by an explanation.

For the purposes of this project, we consider an explanation to be a *transfer of knowledge between agents which allows one agent to understand a belief held by, or an action performed by, the other* [4].

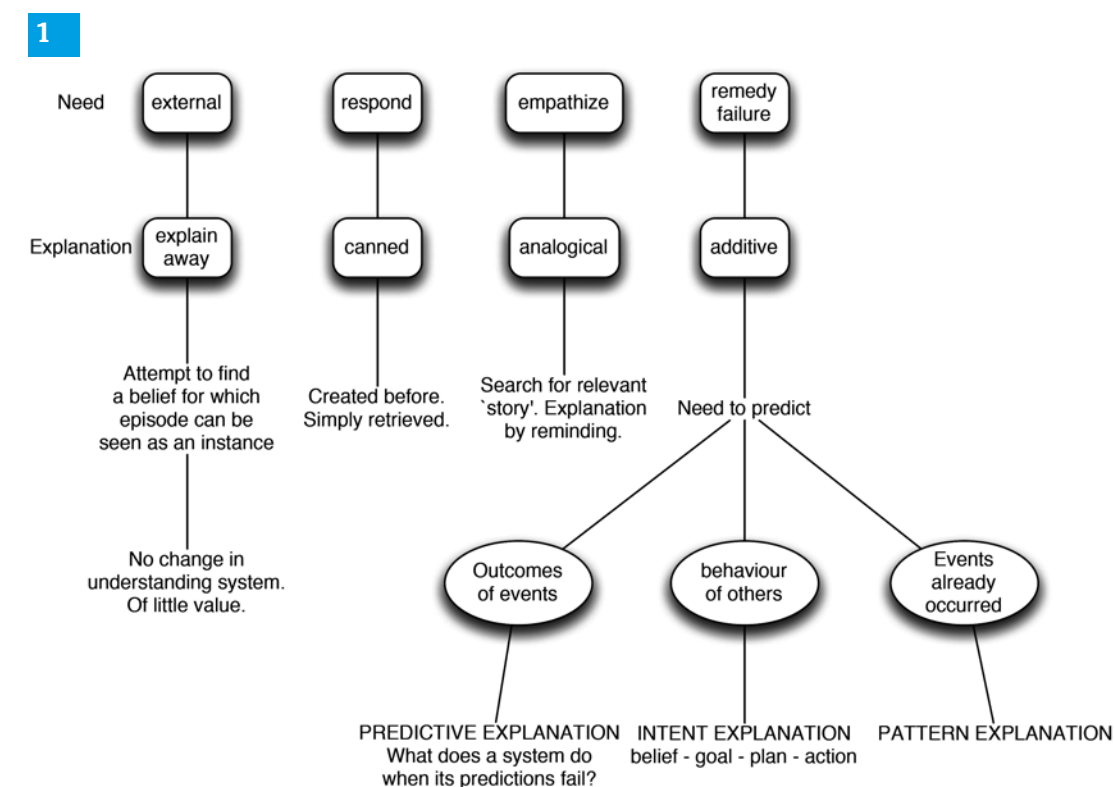
It is not necessary, and indeed not common, that agents in an explanation process have the same *knowledge base* (KB). With the same KB, an explanation can be seen as a trivial transfer of facts and observations. However, with different KBs an explanation process involves a transformation between the two bases. In the case of a human user interacting with an autonomous system it is the latter situation which arises.

It is also worth considering when a user would *require* an explanation from the autonomous system. According to Schank [5] there are four different needs that require an explanation: *external*, *respond*, *empathise* and *remedy failure*. Their respective types of explanation are: *explain away*, *canned*, *analogical* and *additive*.

- **Explain away:** an explanation given to a question whose domain is of little relevance to the explainer. This type of explanation aims to satisfy the question without care of whether the explanation is correct. An example of such explanations are those given on the street by the general public to roaming TV reporters asking arbitrary questions such as; “why do you think the credit crunch happened?”

- **Canned:** canned explanations are those which come from a standard library of explanations already used and deemed correct for certain questions. These are simply retrieved when required.
- **Analogical:** these explanations attempt to connect different domains by drawing similarities between parts of the domains.
- **Additive:** when an observation conflicts with a model of the world, then that observation requires explanation. These are additive in the sense that they add to the model, changing it in some way. This is the most powerful type of explanation and can be subdivided into:
  - predictive: explains a failed prediction
  - intent: explains the intent of another agent
  - pattern: explains past sequences of events.

The need for an explanation and corresponding explanation type are shown schematically in Figure 1. The need for an explanation can arise from any part of an environment or actions of individuals in those environments. In the context of our scenario (the simplified urban search model discussed below), it is worth discussing the explanations that deal with *intent* with more detail.



A graph showing how need for an explanation relates to the explanation type [5].

A generic model for an agent's intent consists of belief, goal, plan and action, defined as follows:

- **Fundamental beliefs:** The beliefs and facts that an agent possesses
- **Goal(s):** The goals that an agent may have. Possibly only one active at a given moment
- **Plan:** The 'high level' action that the agent will perform to achieve the current goal(s)
- **Action:** The atomic action performed at a given moment

## Formal definition of an explanation

We define a generic decision-making system as follows:

*Definition 1:* A decision making system is a mapping

$$\delta : FB \rightarrow A$$

where *FB* are the fundamental beliefs that an agent has about the world that he operates in and *A* is a finite set of actions available to the agent.



