WORKING PAPER 299/2026

# Explainable Decision Support in Multi-Agent AI Systems Using L-Valued Information Flow and Shapley Aggregation

**Purbita Jana**

**MADRAS SCHOOL OF ECONOMICS**
Gandhi Mandapam Road
Chennai 600 025
India

May 2026

# *Explainable Decision Support in Multi-Agent AI Systems Using L-Valued Information Flow and Shapley Aggregation*

**Purbita Jana**
Assistant Professor and
Chair of M.Sc. Data Science Programme,
Madras School of Economics, Chennai, India.
purbita@mse.ac.in

**WORKING PAPER 299/2026**

**May 2026**

**Price : Rs. 35**

**MADRAS SCHOOL OF ECONOMICS**
**Gandhi Mandapam Road**
**Chennai600 025**
**India**

**Phone: 2230 0304/2230 0307/2235 2157**
**Fax: 2235 4847/2235 2155**
**Email : info@mse.ac.in**
**Website: www.mse.ac.in**

# Explainable Decision Support in Multi-Agent AI Systems Using $L$-Valued Information Flow and Shapley Aggregation

Purbita Jana

**Abstract**

Modern AI systems increasingly rely on distributed architectures in which multiple agents, tools, and reasoning modules interact under uncertainty to produce collective decisions. However, existing approaches to decision aggregation and explainability often lack a unified semantic foundation and typically rely on post hoc attribution methods. This paper introduces an explainable multi-agent decision framework based on an L-valued extension of information flow theory. By modeling subsystems as graded semantic classifications connected through structure-preserving information channels, the framework enables coherent aggregation of uncertain information while preserving interpretability. Shapley-value-based attribution is integrated directly into the semantic architecture, yielding intrinsic explanations of subsystem contributions to global decisions. The proposed framework unifies uncertainty modeling, distributed reasoning, and explainable AI within a compositional mathematical structure, with applications to multi-agent systems, tool-augmented language models, and intelligent decision-support systems.

**Keywords:** Explainable AI, Decision Support Systems, Multi-agent systems, Shapley value, Information flow, Fuzzy logic, L-valued systems, Uncertainty modeling, Human-AI decision making
**JEL Codes: C45, C63, C65, D81, D83**

# Acknowledgement

A preliminary version of this work was presented in the MSE Seminar Retreat 2025. The authors are grateful to the participants for their valuable comments and suggestions, which helped improve the present manuscript.

**Purbita Jana**

# 1 Introduction

Contemporary decision-support systems are increasingly driven by distributed artificial intelligence architectures, including multi-agent systems, tool-augmented large language models, and hybrid neuro-symbolic pipelines. In such systems, decisions are no longer produced by a single model but emerge from the interaction of multiple heterogeneous components, each contributing partial, uncertain, and context-dependent information. This shift raises fundamental challenges for decision-making under uncertainty, particularly in ensuring that aggregated decisions are both semantically coherent and explainable to human stakeholders.

Existing approaches to decision aggregation typically rely on probabilistic fusion, heuristic scoring, or optimization-based methods. While effective in specific contexts, these approaches often lack a principled semantic foundation for integrating heterogeneous information sources. Foundational work on information flow and distributed system [1] provides a rigorous account of how information propagates across systems, but assumes binary satisfaction and does not directly address graded uncertainty. In parallel, fuzzy logic and graded reasoning frameworks [5, 2] enable systematic modeling of uncertainty, yet are typically studied independently of distributed information-flow architectures.

At the same time, explainability techniques—most prominently those based on Shapley values—have become central to interpreting AI systems. The unified framework of [6] has established a standard for feature attribution, with further refinements such as [3] and extensions to data valuation in [4]. However, these approaches are typically applied post hoc, treating the system as a black box and producing explanations that are not intrinsically aligned with the underlying decision process.

In parallel, advances in multi-agent systems and AI architectures emphasize modularity and coordination. Classical foundations such as [8] provide a framework for distributed decision-making, while recent developments in tool-augmented language models, including [7] and [9], demonstrate how complex decisions can emerge from coordinated subsystems. Nevertheless, these approaches lack a unified semantic framework that simultaneously captures uncertainty, information flow, and explainability.

This gap between decision formation and decision explanation becomes particularly critical in applications such as risk-sensitive decision-making, intelligent automation, and human–AI collaboration, where transparency and accountability are essential. There is therefore a growing need for frameworks that integrate uncertainty modeling, distributed decision-making, and explainability within a unified and interpretable structure.

In this paper, we propose an explainable multi-agent decision framework based on an L-valued extension of information flow theory. The framework models each subsystem as a graded semantic structure that captures degrees of relevance or confidence, enabling a

natural representation of uncertainty. Information exchange between subsystems is governed by structure-preserving mappings, ensuring that local evaluations are coherently integrated into global decisions without loss of semantic consistency.

A key contribution of this work is the integration of Shapley-value-based attribution directly into the decision architecture. Rather than treating explainability as an external layer, we embed contribution analysis within the process of information aggregation itself. This allows the framework to provide intrinsic explanations of how individual subsystems influence collective decisions, supporting transparency, trust, and accountability in complex AI-driven environments.

The proposed approach is particularly well suited for modern decision-support settings, including multi-agent coordination, AI-assisted tool selection, and modular decision pipelines. Through representative scenarios, we demonstrate how distributed evidence can be aggregated and explained in a principled and interpretable manner.

# 2 Related Work

This work lies at the intersection of information flow theory, fuzzy logic and uncertainty modeling, explainable artificial intelligence, and multi-agent decision systems. We briefly review each of these areas and highlight the gap addressed by our framework.

## 2.1 Information Flow and Distributed Semantics

Information flow theory, developed by Barwise and Seligman [1], provides a semantic framework for modeling how information propagates across distributed systems via classifications and infomorphisms. Its emphasis on structural relationships rather than statistical dependence makes it particularly suitable for reasoning about modular systems. However, the classical formulation assumes binary satisfaction, which limits its applicability in settings where information is inherently uncertain or graded. Consequently, while information flow theory offers a strong semantic foundation, it does not directly address uncertainty-aware decision-making in modern AI systems.

## 2.2 Fuzzy Logic and Uncertainty Modeling

Fuzzy logic provides a well-established framework for representing graded truth and reasoning under uncertainty. Foundational developments in many-valued logic [5] and subsequent extensions to fuzzy topology and graded consequence [2] enable systematic treatment of partial satisfaction and semantic vagueness. These approaches have been widely

applied in decision-making and approximate reasoning. However, they are typically studied in isolation from distributed information-flow architectures and do not explicitly address how multiple heterogeneous subsystems can be integrated within a unified semantic framework.

## 2.3 Explainable AI and Shapley-Based Attribution

Explainability has become a central concern in modern AI systems. Shapley-value-based methods, particularly the framework introduced in [6], provide a principled approach to attributing importance to features or components. Subsequent work has improved computational aspects [3] and extended the framework to data valuation [4]. Despite their success, these methods are predominantly post hoc and operate independently of the underlying system architecture. As a result, explanations are often detached from the semantic structure governing decision formation.

## 2.4 Multi-Agent Systems and AI Architectures

Multi-agent systems provide a natural paradigm for distributed decision-making, where multiple agents contribute to a collective outcome [8]. Recent advances in AI architectures, including tool-augmented language models such as Toolformer [7] and ReAct [9], further emphasize modularity and coordination across components. While these approaches demonstrate the effectiveness of distributed reasoning, they lack a unified framework that integrates uncertainty, semantic information flow, and explainability.

## 2.5 Research Gap

The above strands of research address complementary aspects of modern decision-support systems: semantic structure (information flow), uncertainty modeling (fuzzy logic), and interpretability (Shapley-based explainability). However, these aspects are typically treated separately. Existing frameworks either lack a semantic foundation for distributed information integration or provide explanations that are external to the decision process.

To bridge this gap, we develop an L-valued extension of information flow theory that enables graded semantic reasoning across distributed subsystems. By integrating Shapley-value-based attribution directly into the information flow architecture, the proposed framework provides an intrinsic and unified approach to uncertainty-aware, explainable decision-making.

# 3 Classical Channels and Classifications

Channel theory, developed by Barwise and Seligman, provides a semantic framework for modeling information flow in distributed systems. Its central insight is that information transfer is governed not by signals alone, but by shared classificatory structure linking tokens and types across systems.

## 3.1 Classifications

**Definition 1** (Classification [1])**.** *A* classification *is a triple*

$$\mathcal{A} = (A, \Sigma_A, \Vdash_A),$$

*where:*

- $A$ *is a set of* tokens*, representing concrete instances or states,*
- $\Sigma_A$ *is a set of* types*, representing properties or predicates,*
- $\Vdash_A \subseteq A \times \Sigma_A$ *is a satisfaction relation, where* $a \Vdash_A \alpha$ *means that token* $a$ *is of type* $\alpha$*.*

The satisfaction relation is *binary*: for each token–type pair, the type either holds or does not hold. Classifications thus provide a semantic representation of information states rather than numerical measurements.

## 3.2 Infomorphisms

Information flow between classifications is mediated by *infomorphisms*, which preserve satisfaction structure across systems.

**Definition 2** (Infomorphism [1])**.** *Given two classifications*

$$\mathcal{A} = (A, \Sigma_A, \Vdash_A), \quad \mathcal{B} = (B, \Sigma_B, \Vdash_B),$$

*an* infomorphism

$$(f_1, f_2) : \mathcal{A} \rightleftarrows \mathcal{B}$$

*consists of functions*

$$f_1 : \Sigma_A \to \Sigma_B, \qquad f_2 : B \to A,$$

*such that for all* $b \in B$ *and* $\alpha \in \Sigma_A$*,*

$$b \Vdash_B f_1(\alpha) \quad \textit{iff} \quad f_2(b) \Vdash_A \alpha.$$

Infomorphisms ensure that information is preserved under translation between local vocabularies and contexts. The direction of information flow is from tokens to types, while semantic constraints flow contravariantly.

### 3.3 Channels

A *channel* organizes multiple classifications into a coherent structure supporting distributed inference.

**Definition 3** (Channel)**.** *A* channel *consists of:*

- *a family of classifications* $\{\mathcal{A}_i\}_{i\in I}$*,*
- *a distinguished* core *classification* $\mathcal{C}$*,*
- *infomorphisms* $(f_1^i, f_2^i) : \mathcal{A}_i \rightleftarrows \mathcal{C}$ *for each* $i \in I$*.*

The core serves as a locus of information integration, where evidence from distributed components is combined to support global conclusions. Channels thus model how local information constrains and supports global inference without requiring centralized control or direct communication between all components.

### 3.4 Limitations of the Classical Setting

Classical channel theory assumes a binary satisfaction relation between tokens and types. In contemporary AI systems, such a representation is inadequate: information is often partial, noisy, or uncertain, and enforcing binary satisfaction either discards semantic information or requires ad hoc thresholding.

Moreover, classical channels provide no intrinsic mechanism for attributing responsibility when multiple components jointly support or inhibit a global decision. These limitations motivate the development of graded, $L$-valued classifications and channels, which we introduce in the following section.

## 4 From Classical Channels to $L$-Valued Semantic Information Flow

This section develops a unified semantic framework for information flow, starting from classical channel theory and extending it to graded, explainable, and compositional settings suitable for contemporary AI systems.

## 4.1 $L$-Classifications and $L$-Infomorphisms

To address this limitation, we generalize classifications by allowing satisfaction to take values in a complete lattice $L$.

**Definition 4** (L-Classification)**.** *An $L$-classification is a triple, $\mathbf{A} = (A, \Sigma_A, \Vdash_A)$, consisting of a set of tokens $A$, a set of types $\Sigma_A$ and an $L$-valued fuzzy relation ($\Vdash_A$: $A \times \Sigma_A \to L$) between $A$ and $\Sigma_A$ describes which tokens are of which types in what degree.*

For $L = [0, 1]$, satisfaction values encode degrees of belief, confidence, or semantic relevance. Classical classifications are recovered as the special case $L = \{0, 1\}$.

Information flow between $L$-classifications is captured by graded infomorphisms.

**Definition 5** ($L$-Infomorphism)**.** *An infomorphism between two $L$-clasifications $\mathbf{A} = (A, \Sigma_A, \Vdash_A)$ and $\mathbf{B} = (B, \Sigma_B, \Vdash] a h_B)$ is a pair of functions $(f_1, f_2)$, where $f_1 : \Sigma_A \to \Sigma_B$ and $f_2 : B \to A$ such that,*

$$\Vdash_B (b, f_1(\alpha)) = \Vdash_A (f_2(b), \alpha)$$

*for any $\alpha \in \Sigma_A$ and $b \in B$.*

This ensures that semantic support is preserved exactly under translation between contexts.

In particular if we consider $L = [0, 1]$ then the $L$-infomorphism will be known as fuzzy infomorphism.

**Proposition 1.** *Let $\mathbf{A} = (A, \Sigma_A, \Vdash_A)$ be an $L$-classification. Then*

$$id_{\mathbf{A}}(\equiv (id_{\Sigma_A}, id_A)) : \mathbf{A} \to \mathbf{A}$$

*is an $L$-infomorphism, where $id_{\Sigma_A}$ and $id_A$ are the identity functions on $\Sigma_A$ and $A$ respectively.*

*Proof.* Let $a \in A$ and $\alpha \in \Sigma_A$. Then $id_A(a) = a$, $id_{\Sigma_A}(\alpha) = \alpha$ and

$$\begin{aligned} \Vdash_A (id_A(a), \alpha) &= \Vdash_A (a, \alpha) \\ &= \Vdash_A (a, id_{\Sigma_A}(\alpha)) \end{aligned}$$

Hence $id_{\mathbf{A}}$ is an $L$-infomorphism. □

It is to be noted that $id_{\mathbf{A}}$ will be called as identity $L$-infomorphsm

**Proposition 2.** *Let* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ *are* $L$*-classifications and* $(f_1, f_2) : \mathbf{A} \to \mathbf{B}$, $(g_1, g_2) : \mathbf{B} \to \mathbf{C}$ *are* $L$*-infomorphisms. Then*

$$(g_1, g_2) \circ (f_1, f_2) = (g_1 \circ f_1, f_2 \circ g_2) : \mathbf{A} \to \mathbf{C}$$

*is an* $L$*-infomorphism.*

*Proof.* As $(f_1, f_2) : \mathbf{A} \to \mathbf{B}$ and $(g_1, g_2) : \mathbf{B} \to \mathbf{C}$ are infomorphisms, so we have

$$\Vdash_B (b, f_1(\alpha)) = \Vdash_A (f_2(b), \alpha), \text{ for any } \alpha \in \Sigma_A \text{ and } b \in B$$

and

$$\Vdash_C (c, g_1(\beta)) = \Vdash_B (g_2(c), \beta), \text{ for any } \beta \in \Sigma_B \text{ and } c \in C.$$

$$\begin{aligned} \Vdash_C (c, g_1 \circ f_1(\alpha)) &= \Vdash_C (c, g_1(f_1(\alpha))) \\ &= \Vdash_B (g_2(c), f_1(\alpha)) \\ &= \Vdash_A (f_2(g_2(c), \alpha) \\ &= \Vdash_A (f_2 \circ g_2(c), \alpha) \end{aligned}$$

Hence $(g_1, g_2) \circ (f_1, f_2)$ is an $L$-infomorphism. □

Proposition 2 states that composition of two $L$-infomorphisms defined as above is an $L$-infomorphism.

**Proposition 3.** *If* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ *are* $L$*-classifications and* $(f_1, f_2) : \mathbf{A} \to \mathbf{B}$, $(g_1, g_2) : \mathbf{B} \to \mathbf{C}$, $(h_1, h_2) : \mathbf{C} \to \mathbf{D}$ *are* $L$*-infomorphisms then*

$$(h_1, h_2) \circ ((g_1, g_2) \circ (f_1, f_2)) = ((h_1, h_2) \circ (g_1, g_2)) \circ (f_1, f_2).$$

*Proof.* We have $f_1 : \Sigma_A \to \Sigma_B$, $g_1 : \Sigma_B \to \Sigma_C$, $h_1 : \Sigma_C \to \Sigma_D$, $f_2 : B \to A$, $g_2 : C \to B$, $h_2 : D \to C$ are set functions. So $h_1 \circ (g_1 \circ f_1) = (h_1 \circ g_1) \circ f_1$ and $f_2 \circ (g_2 \circ h_2) = (f_2 \circ g_2) \circ h_2$.

$$\begin{aligned} (h_1, h_2) \circ ((g_1, g_2) \circ (f_1, f_2)) &= (h_1, h_2) \circ ((g_1 \circ f_1, f_2 \circ g_2)) \\ &= (h_1 \circ (g_1 \circ f_1), (f_2 \circ g_2) \circ h_2) \\ &= ((h_1 \circ g_1) \circ f_1, f_2 \circ (g_2 \circ h_2)) \\ &= (h_1 \circ g_1, g_2 \circ h_2) \circ (f_1, f_2) \\ &= ((h_1, h_2) \circ (g_1, g_2)) \circ (f_1, f_2) \end{aligned}$$

Hence $\circ$, defined as in Proposition 2, is associative. □

**Proposition 4.** *If* $\mathbf{A}$, $\mathbf{B}$ *are* $L$*-classifications,* $(f_1, f_2) : \mathbf{A} \to \mathbf{B}$ *is an* $L$*-inforphism,* $id_{\mathbf{A}}$ *and* $id_{\mathbf{B}}$ *are identity* $L$*-infomorphisms then*

$$(f_1, f_2) \circ id_{\mathbf{A}} = (f_1, f_2) \text{ and } id_{\mathbf{B}} \circ (f_1, f_2) = (f_1, f_2).$$

*Proof.* We have $id_{\mathbf{A}} = (id_{\Sigma_A}, id_A) : (A, \Sigma_A, \Vdash_A) \to (A, \Sigma_A, \Vdash_A)$ and $(f_1, f_2) : (A, \Sigma_A, \Vdash_A) \to (B, \Sigma_B, \Vdash_B)$. So $f_1 \circ id_{\Sigma_A} = f_1$ and $id_A \circ f_2 = f_2$ yields,

$$\begin{aligned}(f_1, f_2) \circ id_{\mathbf{A}} &= (f_1, f_2) \circ (id_{\Sigma_A}, id_A) \\ &= (f_1 \circ id_{\Sigma_A}, id_A \circ f_2) \\ &= (f_1, f_2).\end{aligned}$$

Now $id_{\mathbf{B}} = (id_{\Sigma_B}, id_B) : (B, \Sigma_B, \Vdash_B) \to (B, \Sigma_B, \Vdash_B)$. So $id_{\Sigma_B} \circ f_1 = f_1$ and $f_2 \circ id_B = f_2$ yields,

$$\begin{aligned}id_{\mathbf{B}} \circ (f_1, f_2) &= (id_{\Sigma_B}, id_B) \circ (f_1, f_2) \\ &= (id_{\Sigma_B} \circ f_1, f_2 \circ id_B) \\ &= (f_1, f_2).\end{aligned}$$

□

Proposition 1, Proposition 2, Proposition 3 and Proposition 4 together establishes the following theorem.

**Theorem 1.** *$L$-Classifications together with $L$-infomorphisms form a category and is denoted by* $\mathbf{LClass}$.

**Corollary 1.** *Fuzzy classifications together with fuzzy infomorphisms form a category and is denoted by* $\mathbf{FClass}$

## 4.2 Algebraic Operations on $L$-Classifications

$L$-classifications admit natural algebraic constructions supporting modularity and abstraction. These include sums (coproducts) for aggregating independent subsystems, products for joint satisfaction of constraints, reducts for vocabulary restriction, and quotients for semantic abstraction and information compression. Token-level evidence can be lifted to subsystem-level judgments via lattice-compatible aggregation operators.

These operations allow large distributed AI systems to be constructed compositionally while preserving semantic coherence.

### 4.2.1 Sums (Coproducts)

**Definition 6.** *Let* $\{\mathbf{A_i}\}_{i\in I}$ *be a family of* $L$*-classifications. Then* $\coprod_{i\in I}\mathbf{A_i} = (\prod_{i\in I} A_i, \coprod_{i\in I}\Sigma_{A_i}, \Vdash_{\coprod_{i\in I}\mathbf{A_i}})$ *where*

1. $\prod_{i\in I} A_i$ *is the cartesian product of* $\{A_i\}_{i\in I}$*;*
2. $\coprod_{i\in I}\Sigma_{A_i}$ *is the disjoint union of* $\{\Sigma_{A_i}\}_{i\in I}$ *and*
3. $\Vdash_{\coprod_{i\in I} A_i}: (\prod_{i\in I} A_i) \times (\coprod_{i\in I}\Sigma_{A_i}) \to L$ *of* $\coprod_{i\in I}\mathbf{A_i}$ *is defined by,*

$$\Vdash_{\coprod_{i\in I}\mathbf{A_i}} (c,(\alpha,i)) = \Vdash_{\mathbf{A_i}} (c(i),\alpha)$$

*for all* $i \in I$

*is the sum of the family of* $L$*-classifications* $\{\mathbf{A_i}\}_{i\in I}$*.*

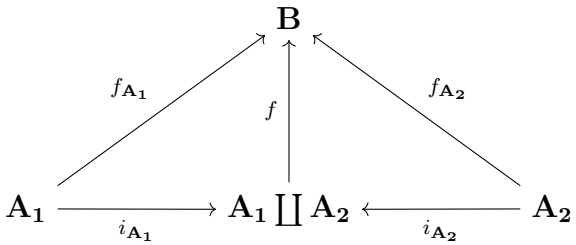
It is possible to show that $\coprod_{i\in I}\mathbf{A_i}$ is the coproduct of the family of $L$-classifications $\{\mathbf{A_i}\}_{i\in I}$ in the category of $L$-classifications.

#### Universal Characterization

**Proposition 5.** *Given infomorphisms* $f_{\mathbf{A_1}}(\equiv (f_{1_{A_1}}, f_{2_{A_1}})): \mathbf{A_1} \to \mathbf{B}$ *and* $f_{\mathbf{A_2}}(\equiv (f_{1_{A_2}}, f_{2_{A_2}})): \mathbf{A_2} \to \mathbf{B}$ *there is a unique infomorphism*

$$f(\equiv (f_1, f_2)): \mathbf{A_1} \coprod \mathbf{A_2} \to \mathbf{B}$$

*such that the following diagram commutes:*




*Proof.* Let us consider $f(\equiv (f_1, f_2)): \mathbf{A_1} \coprod \mathbf{A_2} \to \mathbf{B}$ such that

$$f_1(\alpha,1) = f_{1_{A_1}}(\alpha),\ f_1(\beta,2) = f_{1_{A_2}}(\beta) \text{ and } f_2(b) = (f_{2_{A_1}}(b), f_{2_{A_2}}(b)).$$

To establish the proposition it will be enough to show that $f$ is an infomorphism and it is unique.

$$\begin{aligned}\Vdash_{\mathbf{A_1}((f_{2_{A_1}}(b),f_{2_{A_2}}(b)),\coprod \mathbf{A_2}}\ (\alpha,1)) &= \Vdash_{\mathbf{A_1}}\ (f_{2_{A_1}}(b),\alpha)\\ &= \models_{\mathbf{B}}\ (b, f_{1_{A_1}}(\alpha)).\end{aligned}$$

$$\begin{aligned}\Vdash_{\mathbf{A_1}((f_{2_{A_1}}(b),f_{2_{A_2}}(b)),\coprod \mathbf{A_2}}\ (\beta,2)) &= \Vdash_{\mathbf{A_2}}\ (f_{2_{A_2}}(b),\beta)\\ &= \Vdash_{\mathbf{B}}\ (b, f_{1_{A_2}}(\beta)).\end{aligned}$$

Hence $f$ is an infomorphism.

If possible let there exists $g(\equiv (g_1, g_2))\colon \mathbf{A_1} \coprod \mathbf{A_2} \to \mathbf{B}$ such that

$$g_1 \circ i_{1_{A_1}} = f_{1_{A_1}},\ g_1 \circ i_{1_{A_2}} = f_{1_{A_2}},\ i_{2_{A_1}} \circ g_2 = f_{2_{A_1}},\ i_{2_{A_2}} \circ g_2 = f_{2_{A_2}}.$$

Now $(g_1 \circ i_{1_{A_1}})(\alpha) = g_1(i_{1_{A_1}}(\alpha)) = g_1((\alpha,1))$ and $(f_1 \circ i_{1_{A_1}})(\alpha) = f_1(i_{1_{A_1}}(\alpha)) = f_1((\alpha,1))$. We have $g_1 \circ i_{1_{A_1}} = f_{1_{A_1}} = f_2 \circ i_{1_{A_1}}$. Hence $g_1((\alpha,1)) = f_1((\alpha,1))$ for all $(\alpha,1) \in \Sigma_{A_1} \coprod \Sigma_{A_2}$. Similarly $g_1((\beta,2)) = f_1((\beta,2))$ for all $(\beta,2) \in \Sigma_{A_1} \coprod \Sigma_{A_2}$. Consequently $g_1 = f_1$.

Let $g_2(b) = (c,d)$. We have $(i_{2_{A_1}} \circ f_1)(b) = i_{2_{A_1}}(f_2(b)) = i_{2_{A_1}}(f_{2_{A_1}}(b, f_{2_{A_2}}(b))) = f_{2_{A_1}}(b)$. Hence $i_{2_{A_1}}(g_1(b)) = f_{2_{A_1}}(b)$ and similarly $i_{2_{A_1}}(g_2(b)) = f_{2_{A_2}}(b)$. Now, $f_{2_{A_1}}(b) = i_{2_{A_1}}(g_2(b)) = i_{2_{A_1}}(c,d) = c$ and $f_{2_{A_2}}(b) = i_{2_{A_2}}(g_1(b)) = i_{2_{A_2}}(c,d) = d$. So, $g_2(b) = (f_{2_{A_1}}(b), f_{2_{A_2}}(b)) = f_2(b)$. So $g_1 = f_1$. Hence $(f_1, f_2) = (g_1, g_2)$ consequently $f = g$ and $f$ is unique. □

Sums model *semantic aggregation* of independent information sources. In AI architectures, this corresponds to combining heterogeneous subsystems (e.g., vision, memory, planning) without forcing premature semantic alignment.

### 4.2.2 Products

Given two $L$-classifications

$$\mathcal{A} = (A, \Sigma_A, \Vdash_A), \quad \mathcal{B} = (B, \Sigma_B, \Vdash_B),$$

their *product* is defined as

$$\mathcal{A} \times \mathcal{B} = (A \times B,\ \Sigma_A \times \Sigma_B,\ \Vdash_\times),$$

where

$$\Vdash_\times \big((a,b),(\alpha,\beta)\big) = \Vdash_A (a,\alpha) \wedge \Vdash_B (b,\beta),$$

and $\wedge$ is the meet in $L$.

Products capture *joint satisfaction* of constraints and are useful for modeling synchronized subsystems or conjunctive decision criteria.

### 4.2.3 Reducts and Invariants

**Definition 7** (Invariant)**.** *Given an $L$-classification $\mathbf{A}(\equiv (A, \Sigma_A, \Vdash_{\mathbf{A}}))$, an invariant is a pair $I = (\Sigma, R)$ consisting of a set $\Sigma \subseteq \Sigma_A$ of types of $\mathbf{A}$ and an $L$-valued fuzzy relation $R\colon A \times A \to L$ such that $R(a,b) = \bigwedge_{\alpha\in\Sigma}(\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha))$ and $R(a,a) > 0$ for any $a, b \in A$.*

**Proposition 6.** *$R$ is an $L$-fuzzy equivalence relation.*

*Proof.* $R(a,a) = \bigwedge_{\alpha\in\Sigma}(\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (a,\alpha)) = \bigwedge_{\alpha\in\Sigma} \Vdash_{\mathbf{A}} (a,\alpha)$ and $R(a,b) = \bigwedge_{\alpha\in\Sigma}(\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha))$.

$$\text{Now } \Vdash_{\mathbf{A}} (a,\alpha) \geq \Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha) \text{ for each } \alpha \in \Sigma_A.$$

$$\begin{aligned}\text{Hence } \bigwedge_{\alpha\in\Sigma} \Vdash_{\mathbf{A}} (a,\alpha) &\geq \Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha) \\ &\geq \bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha)).\end{aligned}$$

i.e., $0 < R(a,a) \geq R(a,b)$ for any $a, b \in A$.

$$\begin{aligned}\text{Now } R(a,b) &= \bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha)) \\ &= \bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (b,\alpha) \wedge \Vdash_{\mathbf{A}} (a,\alpha)) \\ &= R(b,a)\end{aligned}$$

$$\begin{aligned}R(a,b) \wedge R(b,c) &= \bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha)) \wedge \bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (b,\alpha) \wedge \Vdash_{\mathbf{A}} (c,\alpha)) \\ &= \bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha) \wedge \Vdash_{\mathbf{A}} (c,\alpha)).\end{aligned}$$

$\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (c,\alpha) \geq \Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha) \wedge \Vdash_{\mathbf{A}} (c,\alpha)$ for all $\alpha \in \Sigma$. So,

$$\begin{aligned}\bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (c,\alpha)) &\geq \Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha) \wedge \Vdash_{\mathbf{A}} (c,\alpha) \\ &\geq \bigwedge_{\alpha\in\Sigma} (\Vdash_{\mathbf{A}} (a,\alpha) \wedge \Vdash_{\mathbf{A}} (b,\alpha) \wedge \Vdash_{\mathbf{A}} (c,\alpha)).\end{aligned}$$

Hence $R(a,c) \geq R(a,b) \wedge R(b,c)$, for all $a, b, c \in A$. □

**Definition 8** (Reduct)**.** *Let $\mathcal{A} = (A, \Sigma, \Vdash)$ be an $L$-classification and let $\Sigma' \subseteq \Sigma$. The* reduct *of $\mathcal{A}$ to $\Sigma'$ is*

$$\mathcal{A} \restriction_{\Sigma'} = (A, \Sigma', \Vdash\restriction), \quad \Vdash\restriction (a, \alpha) = \Vdash (a, \alpha).$$

Reducts support *invariance* under vocabulary restriction and allow subsystems to ignore irrelevant decision types. In AI systems, this corresponds to focusing on task-relevant features or actions while preserving semantic consistency.

### 4.2.4 Quotients and Abstraction

**Definition 9** (Quotient)**.** *Let $I = (\Sigma, R)$ be an invariant on the $L$-classification $\mathbf{A}$. The quotient of $\mathbf{A}$ by $I$, written by $\mathbf{A}/I$, is the $L$-classification with types $\Sigma$, whose tokens are the $R$-equivalence classes of tokens of $\mathbf{A}$ and with $\Vdash_{\mathbf{A}/I} ([a]_R, \alpha) = \Vdash_{\mathbf{A}} (a, \alpha)$.*

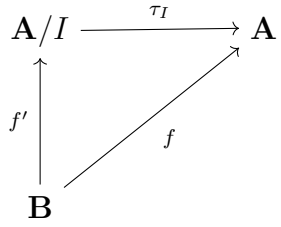
**Definition 10.** *Given an $L$-classification $\mathbf{A}$ and an invariant $I = (\Sigma, R)$ on $\mathbf{A}$, the canonical quotient infomorphism $\tau_I \colon \mathbf{A}/I \to \mathbf{A}$ is the inclusion function on types and on tokens, maps each token of $\mathbf{A}$ to its $R$-equivalence class.*

**Definition 11.** *Given an invariant $I = (\Sigma, R)$ on $\mathbf{A}$, an infomorphism $f (\equiv (f_1, f_2)) \colon \mathbf{B} \to \mathbf{A}$ respects $I$ if*

*i for each $\beta \in \Sigma_B$, $f_1(\beta) \in \Sigma$ and*

*ii if $R(a, b) > 0$, then $f_2(a) = f_2(b)$.*

**Proposition 7.** *Let $I$ be an invariant on $\mathbf{A}$. Given any infomorphism $f \colon \mathbf{B} \to \mathbf{A}$ that respects $I$, there is a unique infomorphism $f' \colon \mathbf{B} \to \mathbf{A}/I$ such that thefollowing diagram commutes.*




*Proof.* Consider $f' (\equiv (f'_1, f'_2)) \colon \mathbf{B} \to \mathbf{A}/I$ such that $f'_1 \colon \Sigma_B \to \Sigma$, $f'_2 \colon A/R \to B$, $f'_1(\beta) = f_1(\beta)$ for all $\beta \in \Sigma_B$ and $f'_2([a]_R) = f_2(a)$ for all $[a]_R \in A/R$. First we need to show that $f'$ is well defined. Let $\beta \in \Sigma_B$, then $f'_1(\beta) \in \Sigma$ and $f_1(\beta) \in \Sigma$, as $f$ respects $I$, for each $\beta \in \Sigma_B$. Hence $f'_1 \colon \Sigma_B \to \Sigma$ defined by $f'_1(\beta) = f_1(\beta)$ for all $\beta \in \Sigma_B$ is

well defined. If $[a]_R = [b]_R$ then $R(a,b) > 0$ for any $[a]_R, [b]_R \in A/R$. As $f$ respects $I$ so $f_2(a) = f_2(b)$ and consequently $f_2'([a]_R) = f_2'([b]_R)$, whenever $[a]_R = [b]_R$. Hence $f_2' \colon A/R \to B$ is well defined. We have,

$$\begin{aligned} \Vdash_{\mathbf{A}/I} ([a]_R, f_1'(\beta)) &= \Vdash_{\mathbf{A}} (a, f_1'(\beta)) \\ &= \Vdash_{\mathbf{A}} (a, f_1(\beta)) \\ &= \Vdash_{\mathbf{B}} (f_2(a), \beta) \\ &= \Vdash_{\mathbf{B}} (f_2(a), \beta) \\ &= \Vdash \mathbf{B}(f_2'([a]_R), \beta). \end{aligned}$$

Hence $f'$ is indeed an infomorphism. □

**Definition 12.** *An infomorphism $f(\equiv (f_1, f_2)) \colon \mathbf{A} \to \mathbf{B}$ is token identical if $A = B$ and $f_2$ is the identity function on $A$. Dually, $f$ is type identical if $\Sigma_A = \Sigma_B$ and $f_1$ is the identity on $\Sigma_A$.*

Quotients implement *semantic abstraction* and information compression. They are particularly useful for privacy preservation, state aggregation, and explainability, where fine-grained distinctions are intentionally collapsed.

### 4.2.5 Aggregation of Tokens

Given an $L$-classification $\mathcal{A} = (A, \Sigma, \Vdash)$ with multiple tokens, a subsystem-level evaluation can be obtained via an aggregation operator

$$\mathsf{Agg} : L^A \to L,$$

such as $\max$, $\min$, or averaging. Define

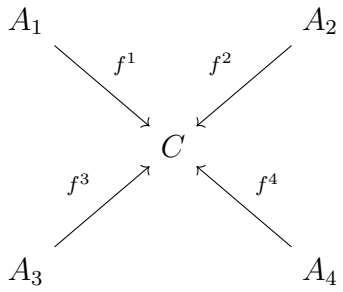
$$\Vdash^{\mathsf{Agg}} (a, \alpha) = \mathsf{Agg}_{x \in A} \Vdash (x, \alpha).$$

This construction allows token-level evidence (e.g., embeddings, retrieved documents, sensor readings) to be lifted to subsystem-level semantic judgments.

## 4.3 $L$-Valued Channels and Aggregation

An *$L$-valued channel* consists of a family of $L$-classifications connected to a core classification via $L$-infomorphisms. The core aggregates information from coalitions of subsystems using a monotone aggregation operator compatible with the lattice structure of $L$.

This provides a principled semantic account of distributed inference under uncertainty, distinct from purely probabilistic or information-theoretic models. Each $\mathcal{A}_i$ models a local AI component (sensor, agent, model, or knowledge base). Information flows through the channel toward the core, supporting global inference or decision-making.




### 4.4 Shapley-Compatible Properties and Explainability

Most existing Shapley-based explanation methods apply contribution attribution post hoc to numerical models [6], treating the system as a black box. While effective for feature attribution, such approaches do not expose the semantic structure governing information flow in distributed systems.

In contrast, our framework integrates Shapley-style attribution directly into the semantics of $L$-valued channels. Coalition values are defined via graded satisfaction at the channel core, ensuring that contribution attribution respects semantic constraints encoded by infomorphisms and is invariant under representation and abstraction.

As a result, explainability emerges as an intrinsic property of the information flow architecture rather than an external interpretability layer, aligning contribution analysis with the logic of distributed reasoning.

## 5 Explainability and Decision-Making

The proposed architecture yields:

- **Semantic coherence:** ensured by $L$-infomorphisms.
- **Uncertainty awareness:** modeled via $[0, 1]$-valued satisfaction.
- **Explainability:** Shapley values quantify contribution and trust.
- **Compositionality:** channels scale to large distributed systems.

# 6 System Architecture: L-Valued Channel-Based Decision System

We propose an $[0,1]$-*valued Channel-Based Decision System (LCDS)* for automated decision-making in contemporary AI architectures, including LLMs with tool augmentation, autonomous robots, and multi-agent systems. The framework is firmly grounded in the information flow theory of [1] and systematically extends it to $[0,1]$-valued classifications, thereby enabling coherent reasoning under uncertainty together with intrinsically modular and interpretable decision mechanisms.

This work deliberately focuses on the unit interval $[0,1]$, viewed as a totally ordered lattice, as it provides a mathematically robust and practically meaningful setting for modeling graded confidence and decision aggregation in real-world systems. At the same time, the framework is designed to admit a substantial generalization beyond this setting. In particular, extending the theory to arbitrary $L$-valued structures over complete lattices constitutes a natural and technically rich next step.

Such a generalization will enable the development of Shapley-value-based attribution mechanisms in non-totally ordered domains, moving beyond the limitations of classical settings. This is expected to significantly enhance the expressive power and applicability of the framework, especially in complex decision environments where uncertainty, preference, and information structures are inherently heterogeneous and only partially ordered.

## 6.1 Distributed Subsystems as $[0,1]$-Classifications

Each functional module of the AI system is modeled as an $[0,1]$-classification

$$\mathcal{A}_i = (A_i, \Sigma_i, \Vdash_i),$$

where $A_i$ is a set of tokens representing internal states (e.g., embeddings, retrieved documents, memories, sensor readings), $\Sigma_i$ is a set of decision-relevant types (e.g., tools or actions), and

$$\Vdash_i \colon A_i \times \Sigma_i \to [0,1]$$

assigns graded semantic relevance.

Typical subsystems include intent detection, semantic retrieval, memory, planning, and safety or policy modules. Each subsystem independently evaluates the suitability of decision types under uncertainty.

## 6.2 Channel Structure and L-Infomorphisms

Subsystems are connected to a central decision module via $[0, 1]$-infomorphisms

$$(f_1^i, f_2^i) : \mathcal{A}_i \rightleftarrows \mathcal{C},$$

which preserve graded satisfaction values across contexts. These mappings align local vocabularies with global decision types and ensure that information flow is semantic rather than heuristic.

The resulting structure forms an *$L$-valued channel*, encoding the regularities that allow information from distributed components to support global inference.

## 6.3 Core Decision Module

The decision module is modeled as a core $[0, 1]$-classification

$$\mathcal{C} = (C, \Sigma_C, \Vdash_C),$$

where $\Sigma_C$ represents candidate decisions (e.g., tool choices) and $\Vdash_C$ aggregates information received from the channel.

For a coalition $S$ of subsystems, aggregation is defined by a lattice-compatible operator:

$$\Vdash_C (c_S, \tau) = \mathsf{Agg}_{i \in S}\big(\Vdash_i (a_i, \tau)\big),$$

where Agg may be instantiated as $\min$, average, or weighted fusion depending on application constraints.

A decision is selected by ranking or thresholding the resulting satisfaction values.

## 6.4 Explainability via Shapley-Based Attribution

To support explainable and trustworthy decisions, each subsystem is treated as a player in a cooperative game. For a fixed decision type $\tau$, define a characteristic function

$$v_\tau : 2^N \to [0, 1], \quad v_\tau(S) = \Vdash_C (c_S, \tau),$$

where $N$ is the set of subsystems.

The Shapley value

$$\phi_i(\tau) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \big(v_\tau(S \cup \{i\}) - v_\tau(S)\big)$$

quantifies the marginal informational contribution of subsystem $i$ to the decision $\tau$. These values provide transparent attribution of influence and support debugging, trust calibration, and responsibility analysis.

### 6.5 Decision Workflow

The end-to-end decision process proceeds as follows:

1. Each subsystem computes local graded relevance scores.
2. Information flows to the core via L-infomorphisms.
3. The core aggregates distributed evidence.
4. A decision is selected based on global satisfaction.
5. Shapley values explain subsystem contributions.

**Summary.** The proposed architecture replaces ad-hoc decision pipelines with a semantically grounded, uncertainty-aware, and explainable system based on $[0, 1]$-valued channels. It scales naturally to multi-tool LLMs, autonomous agents, and distributed AI systems while preserving interpretability and modularity.

## 7 Worked Example: Distributed Decision via a Shapley Channel

All examples in the sequel instantiate the same $L$-valued channel framework, illustrating different choices of aggregation, abstraction, and attribution.

First we illustrate the framework with a simple distributed AI system consisting of three agents providing uncertain evidence toward a global decision.

### 7.1 Local $[0, 1]$-Classifications

Let the system consist of three agents

$$N = \{1, 2, 3\},$$

each modeled as a $[0, 1]$-classification

$$\mathcal{A}_i = (A_i, \Sigma_i, \Vdash_i).$$

- Tokens $A_i$ represent local observations.
- Types $\Sigma_i = \{\mathsf{Threat}\}$ represent a hypothesis.

- Satisfaction values encode belief strength.

Assume the agents assign the following degrees:

$$\Vdash_1 (a_1, \mathsf{Threat}) = 0.6,$$
$$\Vdash_2 (a_2, \mathsf{Threat}) = 0.8,$$
$$\Vdash_3 (a_3, \mathsf{Threat}) = 0.4.$$

## 7.2 Core Classification and Channel

Let the core classification be

$$\mathcal{C} = (C, \Sigma_C, \Vdash_C), \quad \Sigma_C = \{\mathsf{Threat}\}.$$

Each agent is connected to the core via an $[0, 1]$-infomorphism

$$(f_1^i, f_2^i) : \mathcal{A}_i \rightleftarrows \mathcal{C},$$

preserving satisfaction degrees.

We define the core aggregation for a coalition $S \subseteq N$ as

$$\Vdash_C (c_S, \mathsf{Threat}) = \max_{i \in S} \Vdash_i (a_i, \mathsf{Threat}),$$

representing a conservative evidence-fusion strategy.

## 7.3 Characteristic Function

The induced cooperative game for the type Threat is

$$v(S) = \Vdash_C (c_S, \mathsf{Threat}).$$

Thus, 2

- $v(\emptyset) = 0$,
- $v(\{1\}) = 0.6$,
- $v(\{2\}) = 0.8$,
- $v(\{3\}) = 0.4$,
- $v(\{1, 2\}) = 0.8$,
- $v(\{1, 3\}) = 0.6$,
- $v(\{2, 3\}) = 0.8$,
- $v(\{1, 2, 3\}) = 0.8$.

### 7.4 Shapley Value Computation

Using the Shapley value formula, we obtain:

$$\phi_1 = \frac{0.6}{3} + \frac{0}{6} + \frac{0.2}{6} + \frac{0}{3} = 0.233.$$

Similarly,

$$\phi_2 = 0.433, \quad \phi_3 = 0.133.$$

The remaining value corresponds to redundancy shared among agents.

### 7.5 Interpretation

- Agent 2 contributes the most to the global belief.
- Agent 3 contributes marginally due to weaker evidence.
- The channel structure ensures semantic consistency.
- Shapley values provide an explanation-aware attribution of influence.

This example demonstrates how $[0, 1]$-valued channels combine uncertainty, semantic structure, and cooperative attribution to support explainable distributed decision-making.

## 8 Example: Automated Decision-Maker Robot

We model an autonomous robot as a distributed decision-making system whose behavior emerges from structured information flow.

### 8.1 Subsystems as $[0, 1]$-Classifications

Let the robot consist of four subsystems:

$$N = \{V, P, R, T\},$$

corresponding to Vision, Proximity, Risk Analysis, and Task Planning.

Each subsystem is modeled as a $[0, 1]$-classification

$$\mathcal{A}_i = (A_i, \Sigma_i, \Vdash_i).$$

- Tokens represent local system states.

- Types represent decision-relevant propositions.
- Satisfaction values encode confidence or belief.

Let the common decision type be

$$\Sigma_i = \{\mathsf{SafeToAct}\}.$$

Assume the subsystems report:

$$\begin{aligned}
\Vdash_V (v, \mathsf{SafeToAct}) &= 0.7 \quad \text{(vision confidence)},\\
\Vdash_P (p, \mathsf{SafeToAct}) &= 0.9 \quad \text{(proximity sensors)},\\
\Vdash_R (r, \mathsf{SafeToAct}) &= 0.4 \quad \text{(risk assessment)},\\
\Vdash_T (t, \mathsf{SafeToAct}) &= 0.6 \quad \text{(task urgency)}.
\end{aligned}$$

## 8.2 Decision Core and Channel Structure

Let the robot's decision module be the core classification

$$\mathcal{C} = (C, \Sigma_C, \Vdash_C), \quad \Sigma_C = \{\mathsf{ExecuteAction}\}.$$

Each subsystem is connected to the core via an $[0, 1]$-infomorphism

$$(f_1^i, f_2^i) : \mathcal{A}_i \rightleftarrows \mathcal{C},$$

ensuring semantic preservation of confidence values.

The core aggregates information from a coalition $S \subseteq N$ via:

$$\Vdash_C (c_S, \mathsf{ExecuteAction}) = \min_{i \in S} \Vdash_i (\cdot, \mathsf{SafeToAct}),$$

reflecting a safety-critical decision policy.

## 8.3 Characteristic Function

This induces a cooperative game

$$v : 2^N \to [0, 1], \quad v(S) = \Vdash_C (c_S, \mathsf{ExecuteAction}).$$

Key values include: 2

- $v(\emptyset) = 0$,

- $v(\{V\}) = 0.7$,
- $v(\{P\}) = 0.9$,
- $v(\{R\}) = 0.4$,
- $v(\{T\}) = 0.6$,
- $v(\{V, P\}) = 0.7$,
- $v(\{V, R\}) = 0.4$,
- $v(\{V, T\}) = 0.6$,
- $v(\{P, R\}) = 0.4$,
- $v(\{P, T\}) = 0.6$,
- $v(\{R, T\}) = 0.4$,
- $v(\{V, P, R\}) = 0.4$,
- $v(\{V, P, T\}) = 0.6$,
- $v(\{V, R, T\}) = 0.4$,
- $v(\{P, R, T\}) = 0.4$,
- $v(\{V, P, R, T\}) = 0.4$.

## 8.4 Shapley-Based Attribution

The Shapley value for each subsystem is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \big(v(S \cup \{i\}) - v(S)\big).$$

This yields:

$$\begin{aligned} \phi_V &= \frac{0.7}{4} + \frac{-0.2}{12} + \frac{0}{12} + \frac{0}{12} + \frac{0}{12} + \frac{0}{12} + \frac{0}{12} + \frac{0}{4} \\ &= 0.1583. \end{aligned}$$

and similarly,

$$\phi_P = 0.225, \quad \phi_R = -0.09167, \quad \phi_T = 0.10833.$$

### 8.5 Decision and Explanation

Since

$$\Vdash_C (c_N, \mathsf{ExecuteAction}) = 0.4 < \theta,$$

for a safety threshold $\theta = 0.5$, the robot *does not execute the action*.

The Shapley values explain this decision:

- The Risk module contributes the largest negative influence.
- High sensor confidence alone is insufficient.
- The decision emerges from structured information flow, not a single rule.

## 9 Example: LLM Tool Selection via an L-Valued Channel

We model tool selection in a large language model (LLM) as a distributed decision-making process governed by structured information flow.

### 9.1 Subsystems as $[0, 1]$-Classifications

Let the LLM architecture consist of the following subsystems:

$$N = \{I, S, M, C\},$$

where:

- $I$ : Intent detection module,
- $S$ : Semantic similarity module,
- $M$ : Memory / retrieval module,
- $C$ : Contextual constraint module.

Each subsystem is modeled as a $[0, 1]$-classification

$$\mathcal{A}_i = (A_i, \Sigma_i, \Vdash_i).$$

Tokens represent internal representations of the query, while types represent candidate tools.

Let the tool set be:

$$\Sigma_i = \{\mathsf{Search}, \mathsf{Calculator}, \mathsf{Code}, \mathsf{Planner}\}.$$

Each subsystem assigns degrees of suitability:

$$\Vdash_i (a_i, \tau) \in [0, 1],$$

interpreted as confidence that tool $\tau$ is appropriate.

## 9.2 Local Assessments

For a given query, suppose the subsystems provide the following values for the tool Calculator:

$$\begin{aligned}
\Vdash_I (i, \mathsf{Calculator}) &= 0.8 \quad \text{(intent: numerical)},\\
\Vdash_S (s, \mathsf{Calculator}) &= 0.6 \quad \text{(semantic match)},\\
\Vdash_M (m, \mathsf{Calculator}) &= 0.4 \quad \text{(past usage)},\\
\Vdash_C (c, \mathsf{Calculator}) &= 0.9 \quad \text{(context allows)}.
\end{aligned}$$

## 9.3 Core Classification and Channel

The tool-selection module is modeled as the core classification

$$\mathcal{C} = (C, \Sigma_C, \Vdash_C), \quad \Sigma_C = \Sigma_i.$$

Each subsystem connects to the core via an $[0, 1]$-infomorphism

$$(f_1^i, f_2^i) : \mathcal{A}_i \rightleftarrows \mathcal{C},$$

ensuring preservation of graded tool relevance.

The core aggregates coalition information using:

$$\Vdash_C (c_S, \tau) = \frac{1}{|S|} \sum_{i \in S} \Vdash_i (a_i, \tau),$$

representing evidence averaging.

## 9.4 Characteristic Function

Fix a candidate tool $\tau = \mathsf{Calculator}$. Define the characteristic function:

$$v_\tau(S) = \Vdash_C (c_S, \tau), \quad S \subseteq N.$$

Key values include: 2

- $v(\emptyset) = 0$,
- $v(\{I\}) = 0.8$,
- $v(\{S\}) = 0.6$,
- $v(\{M\}) = 0.4$,
- $v(\{C\}) = 0.9$,
- $v(\{I, S\}) = 0.7$,
- $v(\{I, M\}) = 0.6$,
- $v(\{I, C\}) = 0.85$,
- $v(\{S, M\}) = 0.5$,
- $v(\{S, C\}) = 0.75$,
- $v(\{M, C\}) = 0.65$,
- $v(\{I, S, M\}) = 0.6$,
- $v(\{I, S, C\}) = 0.767$,
- $v(\{I, M, C\}) = 0.7$,
- $v(\{S, M, C\}) = 0.633$,
- $v(\{I, S, M, C\}) = 0.675$.

### 9.5 Shapley Value for Tool Attribution

The Shapley value for each subsystem is:

$$\phi_i(\tau) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \big(v_\tau(S \cup \{i\}) - v_\tau(S)\big).$$

This yields:

$$\begin{aligned}\phi_I &= \frac{0.8}{4} + \frac{0.1}{12} + \frac{0.2}{12} + \frac{-0.05}{12} + \frac{0.1}{12} + \frac{0.017}{12} + \frac{0.05}{12} + \frac{0.042}{4} \\ &= 0.24525.\end{aligned}$$

and similarly,

$$\phi_S = 0.12291666667, \quad \phi_M = 0.0005833, \quad \phi_C = 0.30625.$$

### 9.6 Decision and Explanation

Since

$$\Vdash_C (c_N, \mathsf{Calculator}) = 0.675 > \theta,$$

for a selection threshold $\theta = 0.6$, the LLM selects Calculator.

The Shapley values explain the decision:

- Contextual constraints and intent detection dominate the decision.
- Memory plays a minor role.
- Tool selection is justified via structured information flow.

## 10 Complex Example: Multi-Token, Multi-Type LLM Tool Selection

We consider an agentic LLM required to select and sequence tools for a complex query involving data retrieval, numerical analysis, and visualization.

## 10.1 Subsystems and Tokens

Let the distributed subsystems be:

$$N = \{I, S, M, P\},$$

where:

- $I$: Intent and task decomposition,
- $S$: Semantic retrieval,
- $M$: Memory and prior executions,
- $P$: Planning and constraint reasoning.

Each subsystem $\mathcal{A}_i = (A_i, \Sigma_i, \Vdash_i)$ has *multiple tokens*.

$$\begin{aligned} A_I &= \{a_1^{\text{query}}, a_2^{\text{subtask}}\}, \\ A_S &= \{b_1^{\text{GDP}}, b_2^{\text{YearRange}}\}, \\ A_M &= \{c_1^{\text{history}}, c_2^{\text{tool-success}}\}, \\ A_P &= \{d_1^{\text{constraints}}, d_2^{\text{workflow}}\}. \end{aligned}$$

## 10.2 Type Set (Tools and Subtasks)

All subsystems share the type set:

$$\Sigma = \{\mathsf{Search}, \mathsf{Calculator}, \mathsf{CodeExecutor}, \mathsf{Planner}\}.$$

Each type corresponds to a candidate tool.

## 10.3 Local Satisfaction Relations

Subsystems assign graded relevance values for each token–type pair. Selected values are shown below.

| Token | Search | Calc | Code | Plan |
|---|---|---|---|---|
| $a_1^{\text{query}}$ | 0.9 | 0.4 | 0.3 | 0.8 |
| $a_2^{\text{subtask}}$ | 0.6 | 0.7 | 0.5 | 0.9 |
| $b_1^{\text{GDP}}$ | 0.8 | 0.6 | 0.4 | 0.5 |
| $b_2^{\text{YearRange}}$ | 0.7 | 0.5 | 0.4 | 0.4 |
| $c_1^{\text{history}}$ | 0.3 | 0.6 | 0.7 | 0.5 |
| $c_2^{\text{tool-success}}$ | 0.4 | 0.5 | 0.8 | 0.4 |
| $d_1^{\text{constraints}}$ | 0.2 | 0.4 | 0.6 | 0.9 |
| $d_2^{\text{workflow}}$ | 0.3 | 0.5 | 0.7 | 0.8 |

## 10.4 Token Aggregation within Subsystems

Each subsystem aggregates token-level evidence via:

$$\Vdash_i (a, \tau) = \max_{x \in A_i} \Vdash_i (x, \tau),$$

yielding subsystem-level relevance.

## 10.5 Channel Core and Infomorphisms

The LLM tool selector is the core classification:

$$\mathcal{C} = (C, \Sigma, \Vdash_C).$$

Each subsystem is linked to the core via an $[0, 1]$-infomorphism, preserving aggregated satisfaction values.

The core combines subsystem evidence using:

$$\Vdash_C (c_S, \tau) = \frac{1}{|S|} \sum_{i \in S} \Vdash_i (a, \tau).$$

## 10.6 Decision Outcome

For the full coalition $N$, we obtain:

$$\begin{aligned} \Vdash_C (c_N, \textsf{Search}) &= 0.70, \\ \Vdash_C (c_N, \textsf{Calculator}) &= 0.55, \\ \Vdash_C (c_N, \textsf{CodeExecutor}) &= 0.58, \\ \Vdash_C (c_N, \textsf{Planner}) &= 0.78. \end{aligned}$$

Thus, the LLM selects Planner as the first tool, followed by Search and CodeExecutor.

# 11 Experimental Illustration

To demonstrate the practical relevance of the proposed framework, we consider a simplified multi-agent decision-support scenario inspired by AI-assisted tool selection. The objective is to evaluate how distributed components contribute to a final decision under uncertainty and to compare the proposed intrinsic attribution mechanism with standard post hoc approaches.

## 11.1 Setup

We consider a system composed of three agents:

- $A_1$: a reasoning module,
- $A_2$: a retrieval module,
- $A_3$: a tool-execution module.

Each agent produces a graded output in the unit interval $[0, 1]$, representing its confidence or relevance score for a given decision task. These outputs are modeled as L-valued classifications and are aggregated using the proposed information flow mechanism.

For a given instance, suppose the agents produce the following outputs:

$$A_1 = 0.7, \quad A_2 = 0.5, \quad A_3 = 0.6.$$

The global decision score is obtained via a monotone aggregation function consistent with the structure-preserving mappings of the framework. For simplicity, we consider an averaging operator:

$$D = \frac{A_1 + A_2 + A_3}{3} = 0.6.$$

## 11.2 Shapley-Based Attribution

To quantify the contribution of each agent to the final decision, we compute Shapley values based on marginal contributions across all subsets of agents.

Let $v(S)$ denote the aggregated score for a subset $S \subseteq \{A_1, A_2, A_3\}$. The Shapley value $\phi_i$ for agent $A_i$ is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(3 - |S| - 1)!}{3!} \big(v(S \cup \{i\}) - v(S)\big).$$

Using the aggregation rule above, we obtain:

$$\phi_1 \approx 0.233, \quad \phi_2 \approx 0.167, \quad \phi_3 \approx 0.200.$$

These values reflect the relative importance of each agent in determining the final decision.

## 11.3 Comparison with Post Hoc Attribution

We compare the intrinsic attribution produced by the proposed framework with a standard post hoc SHAP-style approximation. In the post hoc setting, attribution is computed after the aggregation step without access to the underlying semantic structure.

While both approaches yield comparable numerical values in this simplified example, a key difference lies in interpretability:

- The proposed method derives contributions directly from the information flow structure, ensuring consistency with the decision process.
- Post hoc methods treat the system as a black box, potentially leading to explanations that are not aligned with internal information propagation.

## 11.4 Discussion

This example illustrates how the proposed framework enables:

- coherent aggregation of distributed, uncertain information,
- principled attribution of decision responsibility,
- alignment between decision formation and explanation.

# 12 Conclusion and Future Work

This paper introduced an explainable multi-agent decision framework based on an L-valued extension of information flow theory. The proposed approach provides a unified semantic foundation for modeling distributed decision-making under uncertainty, where multiple heterogeneous subsystems contribute graded evidence toward a collective outcome. By representing subsystems as L-valued classifications and connecting them through structure-preserving information flow mappings, the framework ensures that local evaluations are coherently integrated into global decisions.

A key feature of the proposed framework is the integration of Shapley-value-based attribution directly within the decision architecture. Unlike conventional post hoc explainability methods, the proposed approach embeds contribution analysis into the process of information aggregation itself. This enables intrinsic, semantically grounded explanations of how individual components influence decision outcomes, thereby enhancing transparency, interpretability, and accountability in complex AI-driven systems.

The framework is sufficiently general to capture a wide range of modern decision-support scenarios, including multi-agent coordination, AI-assisted tool selection, and modular decision pipelines. The illustrative examples demonstrate how uncertainty-aware aggregation and explainability can be achieved in a principled and compositional manner, without relying on ad hoc heuristics or external interpretability layers.

Several directions for future research emerge from this work.

First, the theoretical framework can be extended beyond the unit interval to more general complete lattices, enabling richer representations of uncertainty and preference structures in decision-making environments.

Second, incorporating learning mechanisms into the framework would allow the structure of L-valued channels and aggregation operators to be inferred from data. This would facilitate adaptive decision-support systems that evolve with changing environments and user requirements.

Third, the integration of dynamic and temporal aspects of information flow remains an open problem. Extending the framework to account for sequential decision-making and time-dependent interactions between subsystems would broaden its applicability to real-time and streaming scenarios.

Fourth, further investigation is needed to establish formal properties of the proposed framework, including robustness, stability, and invariance of Shapley-based attribution under transformations such as abstraction and aggregation.

Finally, empirical validation in real-world decision-support applications—such as risk analysis, intelligent automation, and human–AI collaborative systems—would provide valuable insights into the practical effectiveness and scalability of the approach.

Overall, the proposed framework contributes toward bridging the gap between uncertainty modeling, distributed decision-making, and explainable artificial intelligence, offering a principled foundation for the design of transparent and trustworthy decision-support systems.

# References


[1] Jon Barwise and Jerry Seligman. *Information flow*, volume 44 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, Cambridge, 1997. The logic of distributed systems.

[2] Mihir K. Chakraborty and Purbita Jana. Fuzzy topology via fuzzy geometric logic with graded consequence. *Internat. J. Approx. Reason.*, 80:334–347, 2017.

[3] Ian et al. Covert. Improving kernelshap. *AISTATS*, 2021.

[4] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. *ICML*, 2019.

[5] Petr Hájek. *Metamathematics of fuzzy logic*, volume 4 of *Trends in Logic—Studia Logica Library*. Kluwer Academic Publishers, Dordrecht, 1998.

[6] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

[7] Timo et al. Schick. Toolformer: Language models can teach themselves to use tools. *NeurIPS*, 2023.

[8] Michael Wooldridge. *An Introduction to MultiAgent Systems*. Wiley, 2021.

[9] Shunyu et al. Yao. React: Synergizing reasoning and acting in language models. *ICLR*, 2023.

# *MSE Monographs*

* Monograph 36/2017
Underlying Drivers of India's Potential Growth
*C.Rangarajan and D.K. Srivastava*

* Monograph 37/2018
India: The Need for Good Macro Policies *(4th Dr. Raja J. Chelliah Memorial Lecture)*
*Ashok K. Lahiri*

* Monograph 38/2018
Finances of Tamil Nadu Government
*K R Shanmugam*

* Monograph 39/2018
Growth Dynamics of Tamil Nadu Economy
*K R Shanmugam*

* Monograph 40/2018
Goods and Services Tax: Revenue Implications and RNR for Tamil Nadu
*D.K. Srivastava, K.R. Shanmugam*

* Monograph 41/2018
Medium Term Macro Econometric Model of the Indian Economy
*D.K. Srivastava, K.R. Shanmugam*

* Monograph 42/2018
A Macro-Econometric Model of the Indian Economy Based on Quarterly Data
*D.K. Srivastava*

* Monograph 43/2019
The Evolving GST
*Indira Rajaraman*

* Monograph 44/2025 Landscape Analysis of the Labour Market of the Freight Logistics Sector in India
*Gopal Krishna Roy, Brinda Viswanathan, Ashrita. B, Madhuritha Murali and Mohit Sharma*

* Monograph 45/2025 The Fisc and India's Energy Transition
*Laveesh Bhandari*

## *MSE Working Papers*
## Recent Issues

* Working Paper 290/2025
Cournot Equilibrium at the Limit
Naveen Srinivasan, Vijay Adithya C and Poornapushkala Narayanan

* Working Paper 291/2025
Bribing and Vulnerability of the Informal Sector in India
Devlina and Santosh Kumar Sahu

* Working Paper 292/2026

Patent Valuation under Fragile Institutional Enforcement: A Continuous-Time Markov Approach
Srikanth Pai, Akila Hariharan, and Naveen Srinivasan

* Working Paper 293/2026

Early Detection of ESG Policy Violations Using Machine Learning Techniques
Gautami Parate and Arpita Choudhary

* Working Paper 294/2026

Stock Market Reactions to COP26 and Climate Change Exposures of Indian Firms
Saumitra N Bhaduri, Ekta Selarka and Alankrti Aggrwal

* Working Paper 295/2026

European Union Regulations in Indian Tyre Industry
Ekta Selarka and Subrata Sarkar

* Working Paper 296/2026

Women Director Networks and Corporate Social Responsibility of Indian firms
Kavitha Nambiar and Ekta Selarka

* Working Paper 297/2026
Does Perception Matter? The Role of Monetary Policy Uncertainty in Policy Transmission
Aariya Sen

* Working Paper 298/2026

Reimagining Gender Budgeting Framework in India: Linking Fiscal Outlays and Gendered Outcomes
N R Bhanumurthy, Bhabesh Hazarika and Aritri Chakravarty

---

* Working papers are downloadable from MSE website http://www.mse.ac.in