Reasoning Alignment for Agentic AI: Argumentation, Belief Revision, and Dialogue

Tjitze Rienstra
Maastricht University, Maastricht, The Netherlands
t.rienstra@maastrichtuniversity.nl

Leendert van der Torre
University of Luxembourg, Esch-sur-Alzette, Luxembourg and Zhejiang University, China
leon.vandertorre@uni.lu

Liuwen Yu
University of Luxembourg, Esch-sur-Alzette, Luxembourg
liuwen.yu@uni.lu

Abstract

Agentic AI—deployed as technical systems that perceive, decide, and act via tools—faces requirements of safety, accountability, controlled adaptivity, and compositionality. We develop *Reasoning Alignment Diagrams* (RADs), commutative reasoning representations that align a source specification with an argumentation-based explanation route. As illustrative examples, we first show that *full-meet belief base revision* admits an exact representation within *base argumentation* via a restricted-attack construction: the revised base equals the intersection of the premises appearing in all stable extensions of the modified framework. This yields a RAD from input to sanctioned output that doubles as an explanation engine. We then compose the "listen" (revision) and "assert" (inference) RADs to model dialogue among agents, enabling explainable and auditable autonomy. Although our results are entirely symbolic, the RAD template can serve as a specification layer even when other components are opaque or learned. The approach realizes core themes of Gabbay's programme—logic as a toolbox, combining logics, and argumentation as a host formalism—and supports a principle-based analysis of correctness, transparency, and modularity.

Vol. 12 No. 6 2025
Journal of Applied Logics — IfCoLog Journal of Logics and their Applications

1 Dov's vision and approach

Dov Gabbay's research programme treats logic not as a single monolithic calculus but as an *engineerable repertoire of mechanisms* that we assemble to model reasoning in context. Three leitmotifs run through his work and give us a methodological compass:

Logic as a toolbox. Beginning with *Labelled Deductive Systems* (LDS[46]), Gabbay advocates building logics from reusable components—labels for time, agents, resources, priorities; modalities for knowledge and obligation; non-monotonic rules and preference orderings: chosen to fit an application. The point is methodological pluralism with engineering discipline: pick the right tools, and make the choices explicit.

Combining logics. The toolbox needs assembly rules. In *Fibring Logics* and related work [45, 47, 48], Gabbay and collaborators developed operations (fibring, fusion, possible-translations) and meta-results that tell us when properties (soundness, completeness, decidability, interpolation) are preserved or lost when we cut and paste reasoning systems. This provides the algebra of composition required to scale from local components to full systems.

Argumentation as the logic for our century. Gabbay has consistently argued that the centre of gravity in logic must shift from static consequence to *interactive, defeasible, and explainable* reasoning. His contributions to abstract and structured argumentation—e.g., equational/numerical semantics for argumentation networks and higher-order attacks—frame argumentation as the *host formalism* where heterogeneous mechanisms live together and where explanations are native [10, 51].

Two further threads reinforce this picture. First, *temporal and executable views of logic* (e.g., MetateM) recast formulas as rules that drive processes over time; second, *reactive Kripke semantics* let models change as evaluation proceeds [50]. Together, they emphasize dynamics and interaction—exactly what argumentation systems aim to capture.

From this programme we distill five working principles that guide our paper.

P1 (Mechanisms, not monoliths). Choose and expose the representational and inferential devices appropriate to the task (labels, priorities, numerical updates, etc.).

P2 (Modularity and composition). Combine mechanisms using disciplined operations with meta-theoretic guarantees (fibring/fusion/translation), and state what is preserved.

P3 (Dynamics). Treat reasoning as a process that updates states (bases, networks, labels) rather than a once-for-all closure.

P4 (Commutation as explanation). When two perspectives on the same reasoning task coexist (e.g., direct belief change vs. argumentation), require a *commutative diagram*: different routes yield the same outcome. The commuting rectangle provides both a specification and an explanation.

P5 (Transparency). Prefer formalisms that surface *why* an outcome holds. Argumentation earns its keep by turning outcomes into attack/defence structures and (in equational treatments) explicit numerical/functional dependencies.

Our paper instantiates this programme at three levels.

Toolbox to construction. We take two standard "tools" from the KR toolbox—*belief base revision* (AGM-style) and *base argumentation*—and make the design choice that matters for our purpose: in the argumentation framework over $K \cup \{\varphi\}$, we restrict *attacks originating from* $K \setminus \{\varphi\}$. This single, explicit mechanism plays the role of contraction by $\neg\varphi$. The rest is cleanly delegated to Dung-style semantics.

Combining logics to commutation. Rather than embedding one formalism into the other ad hoc, we enforce *commutation* between two routes from input (K, φ) to output $K * \varphi$: revise directly by full-meet, or construct the modified base-AF and compute stable extensions, then extract conclusions. The equality of outcomes is our main theorem. In Gabbay's terms, this is a preservation result: our combination of mechanisms (base arguments + restricted attacks + stable semantics) *preserves* the specification given by full-meet revision.

Argumentation as host for interaction. Argumentation is not merely a target representation; it is the *operating system* in which different reasoning components—revision, inference, explanation—cohabit. We therefore compose two commuting diagrams to model dialogue: *assert* (argumentation-as-inference) and *listen* (revision). This realizes Gabbay's "logic for the 21st century" stance: logic that is interactive (speech acts as moves), defeasible (bases may conflict), dynamic (states evolve), and explainable (attacks/defences or equations justify outcomes) [51, 10].

In contemporary deployments of agentic AI—systems that perceive, decide, and act in tool-rich environments—the need for disciplined composition, dynamics, and explanation resurfaces as an engineering requirement. We instantiate Gabbay's toolbox and combination principles in a reusable pattern we call *Reasoning Alignment Diagrams* (RADs): commutative reasoning representations that align a source specification with an argumentation-based explanation path.

In what follows we first introduce RADs as a general template and discuss the role of this concept in existing applications of Dung's model of abstract argumentation. We then introduce *revision-as-argumentation* RADs to demonstrate that this role extends to belief revision as well. We prove an exact alignment result for revision-as-argumentation RADs (full-meet base revision via restricted attacks in base argumentation), and finally compose a "listen" (revision) RAD with an "assert" (inference) RAD to model dialogue—thus operationalising Gabbay's view of argumentation as the host formalism for interactive, dynamic, and explainable reasoning.

2 Reasoning Alignment Diagrams

In this section we introduce the notion of *reasoning alignment diagram* (RAD) as an architectural pattern for aligning different forms of reasoning. In simple terms, a RAD is a commutative diagram that shows how different forms of reasoning align in the sense that different 'routes' from input to output lead to the same result. As a general concept, a RAD may apply to different types of input and output, and each route between input and output serves a distinct purpose. For instance, the input may be a knowledge base, a state description, a query, a learned model or a combination thereof, while the output may be a set of sanctioned consequences, a prediction, an updated state description, and so on.

A classical example of a RAD arises in logic from the distinction between *semantic entailment* $\models$, typically defined in terms of truth preservation across models, and *proof-theoretic inference* $\vdash$, typically defined in terms of syntactic derivations (see Figure 1). Here, the purpose of the proof-theoretic route is to provide a syntactic, human-interpretable and typically mechanisable procedure that aligns with the semantic definition of entailment. In general, the strongest form of alignment in a RAD is where the outputs of the different routes are equivalent, meaning that the different routes differ only in *how* they arrive at this output. We call this *exact alignment*. In the semantic/proof-theoretic RAD, exact alignment amounts to soundness and completeness of the proof-theoretic inference route. Note that there are other ways to think about alignment, such as *partial alignment*, for instance if an

inference procedure is sound but not complete, and *approximate alignment*, which may be defined in terms of a distance measure on outcomes or probabilistically.

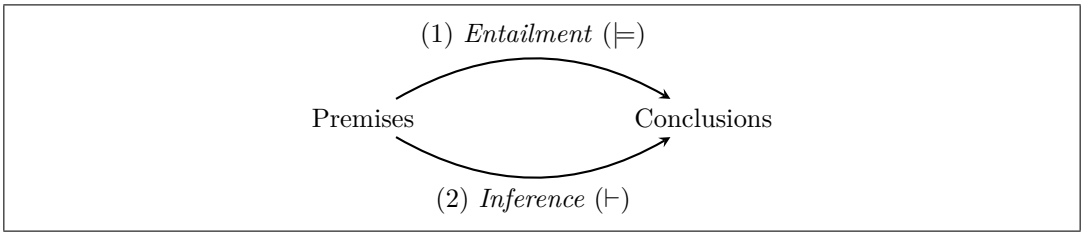

Figure 1: The entailment-as-inference RAD, showing how proof-theoretic inference (route 2) provides an alternative to semantic entailment (route 1).

Another application of the RAD concept is that one route provides a normative justification while another route provides explanations, i.e., one route 'explains' the other. Such RADs may involve more than two arrows to encode translation steps that are needed to obtain explanations. This aligns with how Dung's abstract argumentation is used, which we use in the remainder to develop the concept of RAD. We will discuss three examples of argumentation-based RADs. The first two are *inference-as-argumentation* and *argumentation-as-discussion*, and we discuss prior work that fit these two RAD concepts. The third, which we call *revision-as-argumentation*, is a completely new kind of RAD, where argumentation provides a means to explain the process of *base revision*, a specific form of belief change. In Section 3 we formalise this concept and provide a proof of exact alignment.

The RAD examples that we consider in this section deal with the alignment of different forms of symbolic reasoning. However, as a general concept, it can also be used to align subsymbolic approaches or a combination of symbolic and subsymbolic approaches. Some examples of such RADs will be discussed in Sections 4 and 5.

2.1 Inference-as-Argumentation

An *abstract argumentation framework* consists of a set of arguments together with a binary attack relation [35]. Arguments are treated abstractly, without internal structure, and the attack relation between arguments, where one argument attacks another if the former counts as a counterargument to the latter, determines whether a set of arguments is collectively acceptable. These sets, called *extensions*, represent coherent positions that can be adopted in light of conflicting information. Different criteria can be used to determine the extensions of an argumentation framework, and these criteria define a so-called semantics. Under this view, inference is defined relative to the extensions of the argumentation framework, for example by requiring

that an argument appears in all (*skeptical acceptance*) or some (*credulous acceptance*) extensions under a given semantics.

Dung showed that various forms of defeasible inference can be represented within his theory of abstract argumentation. This elevates the abstract theory into a theory of *structured argumentation*, where an argumentation framework consists of arguments with structured content (e.g. premises, conclusion, and rules that derive the latter from the former) and attacks are induced by conflicts between arguments. This approach is depicted by the RAD shown in Figure 2 [71]. This diagram relates two approaches to deriving conclusions from a knowledge base. The first (arrow 1) is the 'canonical' route defined by the defeasible inference formalism. The other (arrows 2, 3 and 4) is the structured argumentation route. It starts with the translation (arrow 2) of the knowledge base into a structured argumentation framework. Then a semantics (arrow 3) determines the extensions, followed by the extraction of the conclusions of the arguments in the extensions (arrow 4).

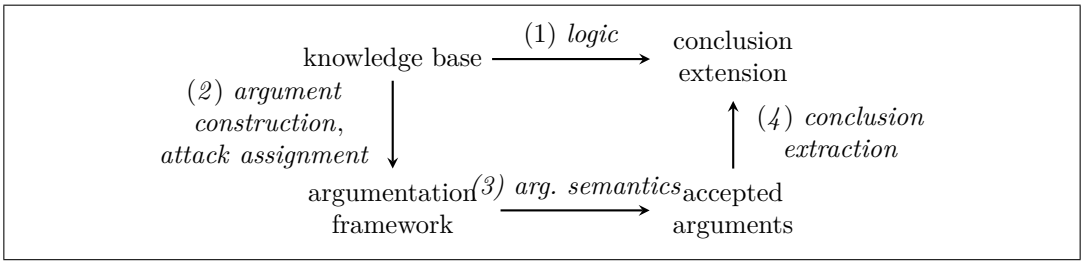

Figure 2: The Inference-as-Argumentation RAD, where the purpose of the argumentation route (arrows 2-3-4) is to explain inferences performed by the source formalism (arrow 1).

Dung's approach can be understood as providing a general framework for *Inference-as-Argumentation* in the sense that the RAD shown in Figure 2 applies to different source formalisms (arrow 1) and corresponding translations (arrows 2 and 4). Dung showed that this scheme applies to Reiter's default logic, Pollock's defeasible reasoning, and logic programming with negation as failure. Others have shown that the approach applies to other forms of inference, such as reasoning with inconsistent knowledge bases using maximally consistent subsets [17, 28]. All of these can be represented as structured argumentation such that the conclusions arrived at using both routes (1 and 2-3-4) coincide. In addition, Dung presented a game-theoretic application of his theory, where arrow 1 corresponds to solving an instance of the stable marriage problem, while arrows 2-3-4 provide an equivalent solution in terms of extensions of an argumentation framework.

Dung's approach can be understood as an *abstraction* in the sense that computing

extensions under a given semantics (arrow 3 in Figure 2) does not depend on the source formalism, or on the content of the constructed arguments. This abstraction is part of its power, as seemingly different forms of inference share the same underlying abstract model of argumentation, a fact that has generated various new insights.

Beyond abstraction, Dung's theory of argumentation provides *explanations*, that is, the derivation of conclusions via an argumentation framework and semantics (arrows 2-3-4) provides a way of *explaining* inferences in the source formalism (arrow 1) as a form of argumentation. An explanation for a conclusion is given by the arguments that justify the conclusion, while the extensions in which these arguments appear demonstrate that these arguments are part of a coherent position. The link between argumentation and explanation has also been investigated in the social sciences, where argumentation and explanation are seen as two closely related human activities [5]. This perspective underpins the role of Dung's model in the wider field of Explainable AI, where argumentation frameworks are used to make reasoning in complex systems more transparent [30, 66].

2.2 Argumentation-as-Discussion

The RAD perspective can also be used to model inference as argumentation at further levels of abstraction. Consider the problem of determining the extensions of an argumentation framework under a given semantics (arrow 3 in Figure 2). This problem has been reformulated in terms of *two-player discussion games*. Such discussion games have been defined to capture acceptance under a number of commonly used argumentation semantics (see [22] for an overview). These are two-player discussion games where the players are typically referred to as *proponent* and *opponent*. Starting from an initial argument put forward by the proponent, the proponent and opponent take turns attacking each other's arguments, with the proponent seeking to defend the claim. The exchange of arguments follows a fixed set of rules, defined in such a way that the existence of a winning strategy for the proponent proves that the initial argument is credulously or skeptically accepted. We can represent this concept with the *Argumentation-as-Discussion* RAD shown in Figure 3. In this RAD, arrow 1 represents the canonical way of determining the extensions of an argumentation framework under a given semantics. The other route (arrows 2-3-4) represents the computation of the extensions using a discussion game (we assume here that winning strategies correspond to extensions). Note that arrow 1 in Figure 3 corresponds to arrow 3 in Figure 2. Combining the RADs in Figure 3 and 2 provides a multi-layer explanatory perspective on inference, with the first layer providing explanations in terms of the extensions of argumentation frameworks, and the second layer in terms of a two-player discussion that establishes the acceptance

of arguments.

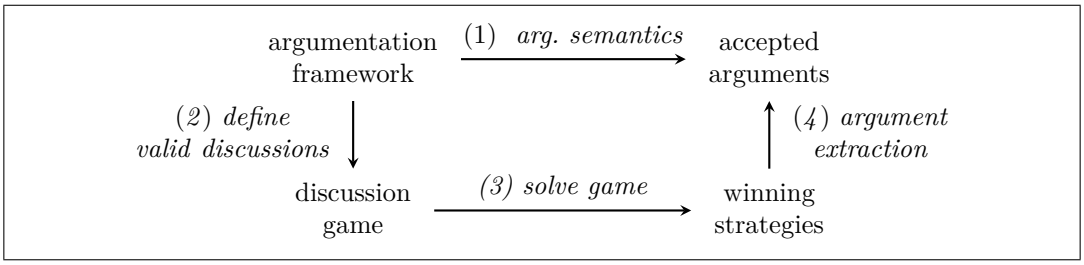

Figure 3: The Argumentation-as-Discussion RAD, showing how discussion games (arrows 2-3-4) explain argument acceptability (arrow 1).

2.3 Revision-As-Argumentation

Belief revision studies how a rational agent should update its beliefs in light of new, possibly conflicting, information [2, 31, 41]. The connection between argumentation and belief revision has been widely studied in the literature, grounded in the view that argumentation is an inherently dynamic process. We discuss work that connect argumentation and belief revision in the related work section. Surprisingly, however, the direct representation of belief revision itself in structured argumentation is largely unexplored. This representation is depicted in the RAD shown in Figure 4. The figure depicts the representation of a specific form of revision (full-meet base revision) as a specific form of structured argumentation (base argumentation), the details of which will be presented in the next section. One route (arrow 1) represents the canonical way of performing full-meet revision: given a belief base K (i.e., a finite, consistent set of formulas) and input ϕ, the revised belief base $K * \phi$ is obtained, in the full-meet case, by contracting $\neg\phi$ and then adding ϕ, where contraction by $\neg\phi$ amounts to taking the intersection of the maximal subsets of K that do not entail $\neg\phi$. The other route (arrows 2-3-4) provides an alternative argumentation route: constructing a base argumentation framework over $K \cup \{\phi\}$, (modifying attacks to encode the contraction step) (arrow 2), computing the extensions (arrow 3), and extracting the revised base (arrow 4). Like in the case of inference-as-argumentation, the argumentation route provides a way to explain the revision process. In the next section we formalise this approach, including the exact alignment of the revision and argumentation routes, thereby demonstrating the possibility of revision-as-argumentation.

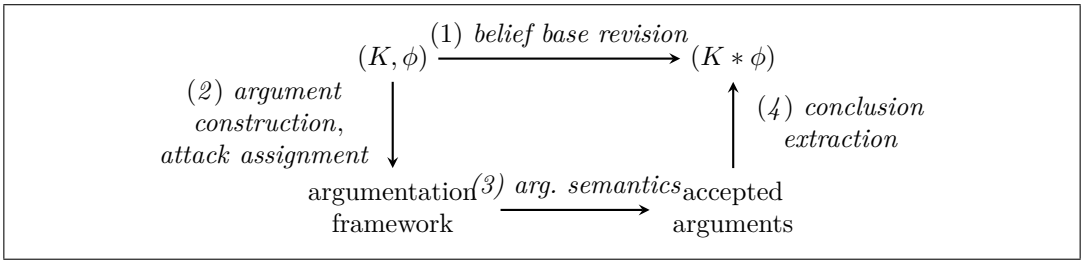

Figure 4: The Revision-as-Argumentation RAD

3 Revision-as-Argumentation Formalised

We now formally develop the *revision-as-argumentation* RAD introduced in the previous section. Like formal argumentation, belief revision [2, 31, 41] is one of the major branches of knowledge representation and reasoning. Connections between formal argumentation and belief revision have been studied from various perspectives in the field of *dynamics of argumentation* [20]. We discuss some of this work in the related work section. However, and perhaps surprisingly, there have been no results to date showing how belief revision itself can be represented as a form of structured argumentation. We believe that the absence of representation results for belief revision as a form of formal argumentation indicates that a general result may be difficult to obtain. Instead, a more promising approach is to seek results connecting specific types of belief revision with particular forms of structured argumentation, which is precisely what we do in this section.

Building on the observation that formal argumentation often operates with sets of formulas that are not deductively closed, we focus on revising such sets: a process referred to as *base revision*. In base revision, a belief base K, a consistent and finite set of formulas, is revised with a new piece of information ϕ, which may be inconsistent with K, to obtain a revised belief base $K * \phi$. Various constructive [56] and postulate-based characterizations [2] of the revision operator $*$ have been proposed. We focus on the simplest form of base revision: *full-meet* base revision [52]. As the formalism for structured argumentation, we adopt *base argumentation*, a recently introduced formalism that is structurally simpler than traditional approaches. Base argumentation represents arguments as minimal consistent subsets of a belief base, with attacks defined by logical contradiction between premises. Despite lacking explicit conclusions, it is equivalent to premise-conclusion argumentation under standard Dung-style semantics. This equivalence is formally established via bisimulation, as shown by Chen et al. [28]. We will show in this section that full-meet

belief base revision can be represented as a kind of base argumentation, thus instantiating the RAD shown in Figure 4. Our aim is to use structured argumentation not merely to reproduce the revised outcome, but to *explain* the revision process by generating an argumentation framework and computing its extensions that yield the same result. Our approach instantiates steps (1, ..., 4) in Figure 4 as follows.

Belief Base Revision: Full-meet base revision of a consistent base K by ϕ is defined via remainder sets: we define $K * \phi = (K - \neg\phi) + \phi$, where contraction $K - \neg\phi$ is the intersection of all remainders of K that do not imply $\neg\phi$.

Argument and attack assignment: Given the belief base $K \cup \{\phi\}$, we construct a base argumentation framework where arguments are minimal consistent subsets of $K \cup \{\phi\}$, and attacks are defined by logical inconsistency. Crucially, we use a *restricted attack relation* $\Rightarrow^{-X}$, which disables attacks originating from arguments entirely within a chosen remainder set X, effectively modeling the contraction step.

Argumentation semantics: We compute the *stable extensions* of this modified base argumentation framework $F(K * \phi) = (\rho(K \cup \{\phi\}), \Rightarrow^{-(K \setminus \{\phi\})})$. Each stable extension corresponds to one maximal consistent subset S of $K \cup \{\phi\}$ that includes ϕ.

Conclusion extraction: The revised base $K * \phi$ is then obtained by intersecting all stable extensions. Thus, the conclusion drawn from the argumentation framework matches the one derived via the belief revision operation.

We first present the necessary definitions of abstract argumentation [35], base argumentation [28] and base revision [52] that we use. We work with the logic $(\mathcal{L}, \vdash)$ where $\mathcal{L}$ is a propositional language defined as usual, and $\vdash$ is the associated consequence relation. Given a set $K \subseteq \mathcal{L}$ we define $Cn(K)$ by $Cn(K) = \{\phi \in \mathcal{L} \mid K \vdash \phi\}$.

3.1 Abstract Argumentation

An abstract argumentation framework is defined as a pair of a set of arguments, and a binary relation representing the attack relationship between arguments [35].

DEFINITION 3.1. *An argumentation framework is a pair* $F = (\mathbf{A}, \Rightarrow)$ *where* $\mathbf{A}$ *is a set of arguments and* $\Rightarrow \subseteq \mathbf{A} \times \mathbf{A}$ *the attack relation.*

Various semantics have been defined as criteria for deciding which sets of arguments are collectively acceptable. For our purpose we will use the *stable semantics.*

DEFINITION 3.2. *Let $F = (\mathbf{A}, \Rightarrow)$ be an argumentation framework, and let $E \subseteq \mathbf{A}$. We say that E is* conflict-free *if there are no $x, y \in E$ such that $(x, y) \in \Rightarrow$, and that E is* stable *if it is conflict-free and attacks every $x \in \mathbf{A} \setminus E$. We denote by $st(F)$ the set of stable extensions of F.*

3.2 Base Argumentation

In most of the instances of inference-as-argumentation (see Figure 2) based on Dung's model of abstract argumentation, an argument is either a recursive structure containing premises, rules and a conclusion, or a pair (Γ, ϕ) where Γ is a set of premises and ϕ is the conclusion. The *base argumentation* formalism was introduced as a simpler and more compact alternative, where a *base argument* is simply a finite and consistent set Γ of formulas. We can think of a base argument Γ as representing the set of all premise-conclusion pairs (Γ, ϕ) such that ϕ follows from Γ. It was shown in [28] that this simpler approach is *extensionally equivalent* to the *deductive argumentation* formalism that uses premise-conclusion pairs [17]: via a simple mapping between base arguments and premise-conclusion pairs, the two induce the same extensions.

Formally, a *base* is a (possibly inconsistent) set $K \subseteq \mathcal{L}$ of formulas that acts as a knowledge base. A common approach to reason with an inconsistent base K is to look at the maximally consistent subsets $MC(K)$ of K defined by

$$MC(K) = \{K' \subseteq K \mid K' \nvdash \bot \text{ and } \forall K'' \text{ s.t. } K' \subset K'' \subseteq K, K'' \vdash \bot\}.$$

Base argumentation provides a way to represent this kind of reasoning as a form of structured argumentation. This corresponds to the RAD depicted in Figure 2, where the source formalism (arrow 1) amounts to computing $MC(K)$ and the argumentation route (arrows 2-3-4) constructing the argumentation framework F_K (arrow 2), computing the stable extensions of F_K (arrow 3), and converting the stable extensions to sets of formulas (arrow 4). The basic definitions for base argumentation are as follows [28].

DEFINITION 3.3. *Let $K \subseteq \mathcal{L}$ be base. A* base argument *of K is a finite set $\Gamma \subseteq K$ that is consistent ($\Gamma \nvdash \bot$) such that there exists no $\Gamma' \subset \Gamma$ with $Cn(\Gamma') = Cn(\Gamma)$. We denote by $\rho(K)$ the set of base arguments of K.*

A base argument Γ attacks another base argument Γ' when Γ entails the negation of some premise in Γ'.

DEFINITION 3.4. *Let $K \subseteq \mathcal{L}$ and let Γ, Γ' be two base arguments of K. Then Γ attacks Γ' iff $\Gamma \vdash \neg\phi$ for some $\phi \in \Gamma'$. We define the attack relation $\rightarrow_K \subseteq \rho(K) \times \rho(K)$ by $\Gamma \rightarrow_K \Gamma'$ iff Γ attacks Γ'.*

We can now define how a base induces a *base argumentation framework*.

DEFINITION 3.5. *The base argumentation framework induced by the set $K \subseteq \mathcal{L}$ is the argumentation framework $F_K = (\rho(K), \rightarrow_K)$.*

EXAMPLE 3.6. *Consider the base $K = \{p, \neg p, q\}$. The base argumentation framework F_K is the pair $(\rho(K), \rightarrow_K)$, where $\rho(K) = \{\emptyset, \{p\}, \{\neg p\}, \{q\}, \{p, q\}, \{\neg p, q\}\}$ and $\rightarrow_K = \{(\{p\}, \{\neg p\}), (\{p\}, \{\neg p, q\}), (\{\neg p\}, \{p\}), (\{\neg p\}, \{p, q\}), (\{p, q\}, \{\neg p\}), (\{p, q\}, \{\neg p, q\}), (\{\neg p, q\}, \{p\}), (\{\neg p, q\}, \{p, q\})\}$.*

The following result (Proposition 3.7) was established by Chen et al. [28]: the stable extensions of the base argumentation framework F_K correspond to the maximal consistent subsets of K. This result establishes the exact alignment of an instance of the RAD depicted in Figure 2 for the case of base argumentation.

PROPOSITION 3.7. *[28] Let $K \subseteq \mathcal{L}$.*

1. *If $S \in MC(K)$ then $\rho(S) \in st((\rho(K), \rightarrow_K))$.*
2. *If $E \in st((\rho(K), \rightarrow_K))$ then $E = \rho(S)$ for some $S \in MC(K)$.*

3.3 Base Revision

Base revision refers to the revision of a base K with a new piece of information ϕ. In the remainder we assume that K is initially consistent, but that K is not necessarily consistent with ϕ. Base revision is modeled by a base revision operator $*$ where $K * \phi$ represents a revision of K with ϕ. Various constructive and postulate-based characterisations for a base revision operator $*$ have been considered (see, e.g., [52]). We present here the constructive definition of the *full-meet revision* operator. First, a *remainder* of K with respect to ϕ is a maximal subset of K that does not imply ϕ [3]:

DEFINITION 3.8 (Remainders)**.** *Let $K \subseteq \mathcal{L}$ and $\phi \in \mathcal{L}$. The set of* remainders *of K with respect to ϕ, denoted $K \perp \phi$, is the set of all subsets $K' \subseteq K$ such that $K' \nvdash \phi$, and there is no K'' with $K' \subset K'' \subseteq K$ and $K'' \nvdash \phi$.*

Revision is defined in terms of *expansion* and *contraction* [2]. Expansion refers to the addition of a belief without checking consistency. Full-meet contraction takes the intersection of all maximal subsets of the original belief set that do not entail the proposition being contracted. Full-meet revision of K with ϕ is defined by the contraction of K by $\neg\phi$ followed by the expansion with ϕ [2].

DEFINITION 3.9. *Let $K \subseteq \mathcal{L}$ and $\phi \in \mathcal{L}$.*

- *The* expansion *of* K *by* ϕ *is defined as:* $K + \phi = K \cup \{\phi\}$.
- *The* full-meet contraction *of* K *by* ϕ *is defined as:* $K - \phi = \bigcap(K \perp \phi)$.
- *The* full-meet revision *of* K *by* ϕ *is defined as:* $K * \phi = (K - \neg\phi) + \phi$.

In what follows we will focus on the operation of full-meet revision as defined above. For further discussion and motivation for these operators, including their characterisation in terms of postulates, we refer the reader to [52].

3.4 Explaining Base Revision Using Base Argumentation

We now show how to model the revision $K * \phi$ as a form of base argumentation. Below we define, given a base K and set $X \subseteq K$, the attack relation $\Rightarrow_X$. In words, given two arguments Γ, Δ the attack $\Gamma \Rightarrow_X \Delta$ holds whenever $\Gamma \rightarrow_K \Delta$ (i.e., Γ attacks Δ according to Definition 3.4) and Γ is not constructed only from elements of X.

DEFINITION 3.10. *Let* $K \subseteq \mathcal{L}$ *and* $X \in MC(K)$. *We define the attack relation* $\Rightarrow_X \subseteq \rho(K) \times \rho(K)$ *by* $\Gamma \Rightarrow_X \Delta$ *iff* $\Gamma \rightarrow_K \Delta$ *and* $\Gamma \not\subseteq X$.

Proposition 3.7 states that, if K is inconsistent then The following lemma states that, if K is inconsistent and X is a maximally consistent subset of K, then the base argumentation framework constructed using the attack relation $\Rightarrow_X$ excludes X.

LEMMA 3.11. *If* $K \subseteq \mathcal{L}$ *is consistent then for every* $X \subseteq K$, $st((\rho(K), \Rightarrow_X)) = \{\rho(K)\}$. *Now suppose* K *is inconsistent and let* $X \in MC(K)$.

1. *If* $S \in MC(K) \setminus \{X\}$ *then* $\rho(S) \in st((\rho(K), \Rightarrow_X))$.
2. *If* $E \in st((\rho(K), \Rightarrow_X))$ *then* $E = \rho(S)$ *for some* $S \in MC(K) \setminus \{X\}$.

Proof. If K is consistent then $\Rightarrow_X = \emptyset$ for every $X \subseteq K$. It then follows that $st((\rho(K), \Rightarrow_X)) = \{\rho(K)\}$. Now assume that K is inconsistent. Let $X \in MC(K)$.

(1) Suppose $S \in MC(K) \setminus \{X\}$. We show that $\rho(S) \in st((\rho(K), \Rightarrow_X))$. We first have that $\rho(S)$ is conflict-free. Otherwise, there exist $\Gamma, \Gamma' \in \rho(S)$, $\Gamma \Rightarrow_X \Gamma'$ and hence $\Gamma \rightarrow_K \Gamma'$. But this implies that S is not consistent, which is false, hence $\rho(S)$ is conflict-free. Now let $\Delta \in \rho(K) \setminus \rho(S)$. Since $S \in MC(K)$, we know $S \cup \Delta$ is inconsistent. Hence there is a $\Gamma \in \rho(S)$ with $Cn(\Gamma) = Cn(S)$ and $\Gamma \rightarrow_K \Delta$. Since $S \neq X$, $\Gamma \not\subseteq X$ and hence $\Gamma \Rightarrow_X \Delta$. Therefore $\rho(S)$ attacks every $\Delta \in \rho(K) \setminus \rho(S)$. It follows that $\rho(S) \in st((\rho(K), \Rightarrow_X))$.

(2) Let $E \in st((\rho(K), \Rightarrow_X))$. Let $S = \cup E$. We will show that (i) $E = \rho(S)$, (ii) $S \in MC(K)$, and (iii) $S \neq X$.

(i) We will show that $E = \rho(S)$. Since K is inconsistent we have that $E \subset \rho(K)$. We first prove that $E = \rho(S)$. Suppose $\Gamma \in \rho(S)$. Suppose for contradiction that $\Gamma \notin E$. Then since E is stable, there is a $\Delta \in E$ s.t. $\Delta \Rightarrow_X \Gamma$. Then $\Delta \vdash \neg\phi$ for some $\phi \in \Gamma$. But since $\Gamma \in \rho(S)$ and $E = \rho(S)$ it follows that there is a $\Gamma' \in E$ such that $\phi \in \Gamma'$. It follows that $\Delta \rightarrow_K \Gamma'$ and (since $\Delta \Rightarrow_X \Gamma$) also $\Delta \Rightarrow_X \Gamma'$. Since $\Delta, \Gamma' \in E$ it follows that E is not conflict-free, which is false. Hence $\Gamma \in E$ and it follows that $E = \rho(S)$.

(ii) We now prove that $S \in MC(K)$. We first prove that S is consistent. Let $\{\Gamma_1, \ldots, \Gamma_n\} = MC(S)$. Since $E = \rho(S)$ it follows that $\Gamma_1, \ldots, \Gamma_n \in E$. If $n = 1$ then $\Gamma_1 = S$ is consistent and we are done. Now suppose for contradiction that $n > 1$. Let $\Delta \in \rho(K) \setminus E$ (existence of Δ follows our assumption that $E \subset \rho(K)$). Stability of E implies that there is a $\Delta' \in E$ such that $\Delta' \Rightarrow_X \Delta$. Since $\Delta' \in \Gamma_i$ for some i it follows that $\Delta' \rightarrow_K \Gamma_j$ for some $i \neq j$. Then, since $\Delta' \Rightarrow_X \Delta$, we also have $\Delta' \Rightarrow_X \Gamma_j$. Since $\Delta, \Gamma_j \in E$ it follows that E is not conflict-free, which is false. Hence S is consistent. Now assume for contradiction that there is consistent S' such that $S \subset S' \subseteq K$. Let $\phi \in S' \setminus S$. Then $\{\phi\} \in \rho(K)$. Since $E = \rho(S)$ it follows that $\{\phi\} \notin E$. Stability of E then implies that for some $\Gamma \in E$, $\Gamma \Rightarrow_X \{\phi\}$ and hence $\Gamma \rightarrow_K \{\phi\}$. This implies that $S \vdash \neg\phi$ which is false since S' is consistent. Hence $S \in MC(K)$.

(iii) We now prove that $S \neq X$. Suppose for contradiction that $S = X$. Let $\Delta \in \rho(K) \setminus E$ (existence of Δ follows our assumption that $E \subset \rho(K)$). Stability of E implies that there is a $\Delta' \in E$ such that $\Delta' \Rightarrow_X \Delta$. But since $\Delta' \subseteq S$ and $S = X$ we have $\Delta' \subseteq X$ and hence $\Delta' \not\Rightarrow_X \Delta$, which is a contradiction. Hence $S \neq X$.

□

We can use this lemma to model full-meet revision of a consistent belief base. Recall that full-meet revision is defined by $K * \phi = (K - \neg\phi) + \phi$. We define the argumentation framework $F_{(K*\phi)}$ by

$$F_{(K*\phi)} = (\rho(K \cup \{\phi\}), \Rightarrow_{(K \setminus \{\phi\})}).$$

The following theorem establishes the link between full-meet base revision $K * \phi$ and the stable extensions of the argumentation framework $F_{(K*\phi)}$.

THEOREM 3.12. *For every consistent $K \subseteq \mathcal{L}$ and $\phi \in \mathcal{L}$:*

$$K * \phi = \bigcap \{\cup_{\Gamma \in E} \Gamma | E \in st(F_{(K*\phi)})\}.$$

Proof. Let $K \subseteq \mathcal{L}$ be a consistent belief base and let $\phi \in \mathcal{L}$ be a consistent formula. We then have

$$\begin{aligned} K * \phi &= (K - \neg\phi) + \phi && (1)\\ &= (\bigcap K\bot\neg\phi) + \phi && (2)\\ &= \{\phi\} \cup \bigcap(K\bot\neg\phi) && (3)\\ &= \bigcap(\{\phi\} \cup K' | K' \in K\bot\neg\phi) && (4)\\ &= \bigcap\{K' \in MC(K \cup \{\phi\}) | \phi \in K'\} && (5)\\ &= \bigcap\{\cup_{\Gamma\in E}\Gamma | E \in st(\rho(K \cup \{\phi\}), \Rightarrow_{K\setminus\{\phi\}})\} && (6)\\ &= \bigcap\{\cup_{\Gamma\in E}\Gamma | E \in st(F_{(K*\phi)})\} && (7) \end{aligned}$$

Justification: In steps (1), (2) and (3) we apply, respectively, the definitions of revision, contraction and expansion. Step (4) follows directly. For (5) we will prove that $(\{\phi\} \cup K' | K' \in K\bot\neg\phi) = \{K' \in MC(K \cup \{\phi\}) | \phi \in K'\}$:

($\subseteq$) Suppose $L \in (\{\phi\} \cup K' | K' \in K\bot\neg\phi)$. Two cases:

Case 1: $\phi \in K$. Then $\forall K' \in K\bot\neg\phi$, $\phi \in K'$. Hence (by our assumption) $L \in K\bot\neg\phi$. Then there is no K' such that $L \subseteq K' \subseteq K$ and $K' \vdash \neg\phi$ and hence (since $\phi \in K'$) no such K' such that $K' \vdash \bot$. This implies that $L \in MC(K \cup \{\phi\})$ and $\phi \in L$.

Case 2: $\phi \notin K$ Then $K \setminus \{\phi\} \in K\bot\neg\phi$. Suppose for contradiction that $L \notin MC(K \cup \{\phi\})$. Then let K' be s.t. $L \subset K' \subseteq K$ s,t. $K' \in MC(K \cup \{\phi\})$. Then $K' \nvdash \neg\phi$ (since $K' \vdash \neg\phi$ would imply $K' \vdash \bot$) and hence $K' \in K\bot\neg\phi$, which is a contradiction.

It follows that $L \in MC(K \cup \{\phi\})$ and $\phi \in K'$.

($\supseteq$) Suppose $L \in \{K' \in MC(K \cup \{\phi\}) | \phi \in K'\}$. Then $L \in MC(K \cup \{\phi\})$ and $\phi \in L$. Suppose $L \notin K\bot\neg\phi$. Then there exists K' with $L \subset K' \subseteq K \cup \{\phi\}$ s.t. $K' \nvdash \neg\phi$. But then K' is a consistent subset of $K \cup \{\phi\}$, contradicting $L \in MC(K \cup \{\phi\})$. It follows that $L \in K\bot\neg\phi$ and hence $L \in (\{\phi\} \cup K' | K' \in K\bot\neg\phi)$.

It follows that $\bigcap(\{\phi\} \cup K' | K' \in K\bot\neg\phi) = \bigcap\{K' \in MC(K \cup \{\phi\}) | \phi \in K'\}$, which proves step (5). Step (6) follows from Lemma 3.11 and step (7) from the definition of $F_{(K*\phi)}$. □

This result establishes a direct correspondence between full-meet belief base revision and structured argumentation, thereby proving the exact alignment of the RAD depicted in Figure 4. We thus show that this form of belief revision can be naturally represented within a structured argumentation framework, which underlines Dov Gabbay's vision of argumentation as a transparent *host formalism* where heterogeneous mechanisms live together [51, 10].

4 Outlook: Multi-RAD Systems and Neuro-symbolic RADs for Agentic AI

In this section we present our vision for how *Reasoning Alignment Diagrams* extend to AI reasoning in the sense of *agentic AI*. At the heart of alignment is making a logical (symbolic) reasoner and a subsymbolic reasoner cohere; our outlook is to develop RADs for *subsymbolic and neurosymbolic* components. In Section 4.1 we state our general vision of *combining RADs* (Multi-RADs) and use the composition of inference-as-argumentation with revision-as-argumentation as an example. In Section 4.2, we use the composition for single-agent decision making. In Section 4.3 we envision a *principle-based* approach for specifying and verifying Multi-RAD systems at their input–output interfaces, independent of internal implementation.

4.1 Multi-RAD for Dialogue

In this section, we take a broader perspective and envision how Inference-as-Argumentation RAD and Revision as Argumentation RAD can be combined for dialogue between agents. The idea is that belief revision and argument generation are essential components of *listening* and *asserting*. This duality allows us to interpret dialogue protocols as the composition of two commutative diagrams. A simple form of dialogue, such as that used in chatbot interaction, can be modeled by speech acts that assert a formula ϕ, thereby informing another agent. This interaction involves belief revision on the receiving side and argument generation on the asserting side. We analyze this kind of dialogue from the perspective of formal argumentation by combining argumentation-as-inference (Figure 2) and argumentation-as-revision (Figure 4), thereby integrating the explanatory structure of base revision with the generative structure of argumentation.

Figure 5 illustrates how revision and inference interact in dialogue. There are three agents, α, β, and γ, each with its own knowledge base: KB_α, KB_β, and KB_γ, respectively. The figure is structured as a temporal and concurrent model: vertical swimlanes represent the internal evolution of each agent's knowledge over time, while horizontal arrows indicate inter-agent communication via speech acts. The process begins with Agent α, who performs argumentation as inference over its knowledge base KB_α. Based on the resulting extension E, α asserts a formula ψ to Agent β. Upon receiving ψ, Agent β listens and revises its knowledge base from KB_β to KB'_β. From this updated base, β constructs a new argumentation framework, infers an extension E, and generates a new assertion δ. In a multi-agent context, an agent may interact with several other agents. Here, Agent γ also asserts δ to Agent β, who integrates it via another revision step from $KB'\beta$ to $KB''\beta$. This update leads to

the generation of a new assertion λ, which is sent to Agent α. Agent α then revises its own knowledge base accordingly, from KB_α to KB'_α.

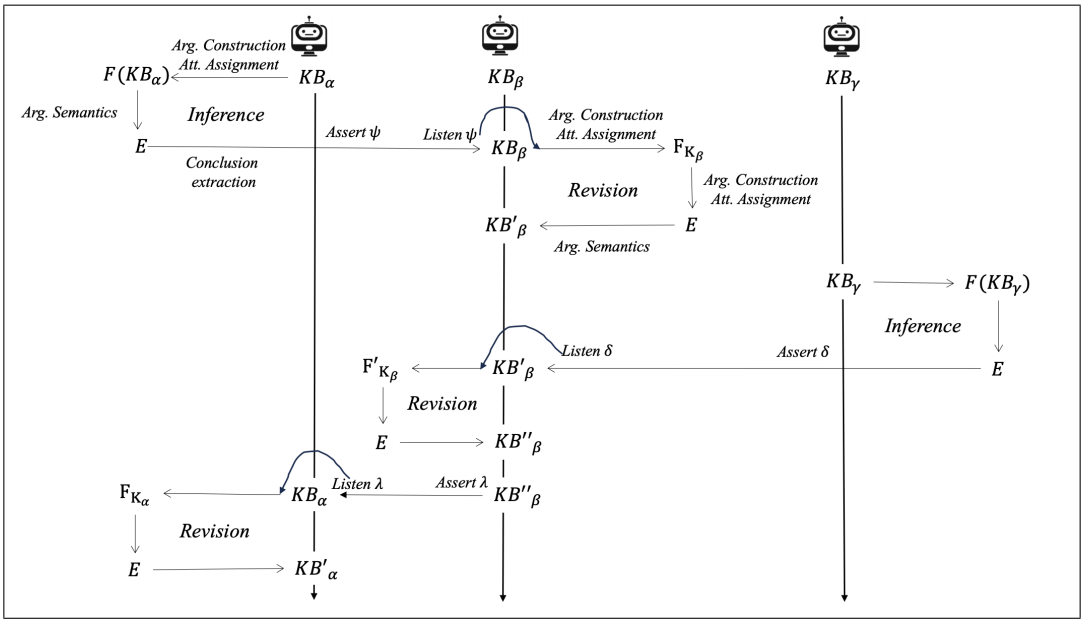

Figure 5: Multi-RAD for dialogue.

4.2 Multi-RAD for Individual Agent Decision Model

In Figure 5, we illustrated combining two RADs in a dialogue context. In this section, we turn to the agent decision model depicted in Figure 6. Whereas Figure 5 considers multiple agents and the evolution of time, and is thus mainly concerned with the synchronization of the agents, Figure 6 examines the interaction from the perspective of a single agent at a fixed point in time.

From the perspective of a single agent α, that agent can either assert something or listen to other agents. If α decides to make an assertion, it must also choose strategically *what* to say. Alternatively, if α is listening and another agent informs it of some proposition ψ, then α has several options: it can revise its knowledge base with the new information, ignore ψ, question the other agents about it by asking for a justification, challenge ψ, and so on. Such agent communication languages based on speech acts have been developed, for example, by FIPA [42].

The strategic decision-making aspect of the dialogue can be implemented using a traditional or a qualitative decision theory [32], as well as techniques from generative

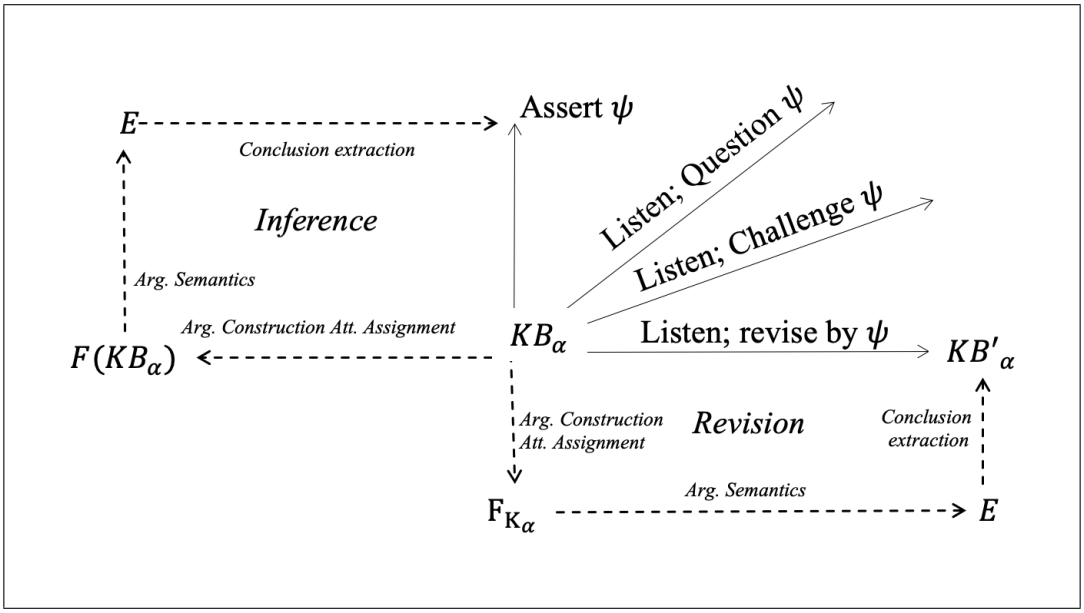

Figure 6: Multi-RAD for individual agent decision

AI. In fact, techniques from neurosymbolic and agentic AI can be adopted. The main question is control: does the symbolic layer or the subsymbolic layer decide?

Moreover, multiple types of dialogues [67] may be distinguished. In persuasion and deliberation dialogues, agents can question or challenge each other's assertions, whereas in information-seeking dialogues, agents primarily ask questions (including requests for clarification). Once an agent is equipped with a range of communicative actions, a meta-level control mechanism is required for deciding when to switch between these actions.

Building on this multi-agent dialogue model, one can investigate a wide range of topics discussed in the multi-agent systems community. For instance, the model could be extended in the direction of hybrid human–machine intelligence, using hybrid human+machine argumentation and notions of cognitive delegation [43] to address questions of autonomy and discretion. Finally, shifting focus from the inter-agent protocol to the viewpoint of a single agent raises broader philosophical questions about agency, which would need to be addressed as this line of research progresses.

4.3 Multi-RAD for Agentic AI

Until now we have shown how individual RADs model inference and revision as argumentation, and how they compose into dialogue and decision-making. Our outlook now moves to apply *Multi-RAD systems* to AI reasoning in the sense of agentic AI.

Agentic AI concerns capabilities such as autonomy, goal pursuit, planning, tool use, and interaction [1]. These capabilities do not presuppose a single implementation style but in hybrid combinations. What matters for us is how the underlying reasoning routes are *aligned* and *composed*. The following example from Gabbay [49] illustrates why multiple kinds of reasoning often co-occur and must be coordinated.

EXAMPLE 4.1 (Untidy room [49])**.** *A mother goes into her teenage daughter's bedroom. Her instant impression is that it is a big mess. There is stuff scattered everywhere. The mother's feeling is that it is not like her daughter to be like this. What happened?*

Conjecture: *The girl may be experiencing boyfriend issues.*

Further Analysis: *The mother notices a collapsed shelf and realizes that the disarray is due to the shelf collapsing under excessive weight which, upon reflection, follows a logical (gravitational) pattern.*

Several types of reasoning are illustrated through this scenario:

Neural network reasoning: *The mess is perceived instantly, similarly to facial recognition by neural networks.*

Nonmonotonic deduction: *The mother deduces from the context and her knowledge that her daughter does not typically live in disarray. Thus, something extraordinary must have happened.*

Abductive reasoning: *She hypothesizes a plausible explanation that her daughter has social-emotional issues, which is common among teenagers.*

Database AI deduction: *A reevaluation leads to the understanding that the mess is due to gravitational effects rather than disorganization on the part of her daughter.*

Pattern recognition: *Someone accustomed to similar patterns may identify the cause as easily as they might recognize a face.*

This example illustrates the need for *hybrid* reasoning, where symbolic and subsymbolic approaches complement one another. By *subsymbolic* reasoning we mean

parameterised function-approximation methods (e.g., neural networks and pattern recognition) that support fast perception and intuitive classification; by *symbolic* reasoning we mean rule- or graph-based methods (e.g., deduction, abduction, database-style re-evaluation) that support structured analysis, logical constraints, and explanations [39]. Neither alone suffices: subsymbolic methods lack transparency, while symbolic methods lack perceptual grounding and flexibility.

To address this, we outline *Multi-RAD systems*: architectures that interconnect several RADs so that each route aligns a source specification with an argumentation-based explanation, and the composed diagram preserves alignment from input to outcome. In practice this means (i) supporting heterogeneity (symbolic and/or subsymbolic components), (ii) ensuring compositionality (routes do not break one another), and (iii) maintaining alignment (each route has a specification and an explanation path that commute).

Multi-RAD systems can embody several conceptualisations of formal argumentation —*inference*, *dialogue*, and *balancing*—which should not be treated in isolation (see [71]). A higher-level metamodel such as A-BDI [70] provides the necessary abstraction, showing how these conceptualisations relate. The next step is a methodology for specifying and verifying such systems, which we introduce below.

4.4 The principle-based approach for multi-RAD systems

To manage the diversity of reasoning methods, we use *principle-based analysis* as a methodology for selecting among existing methods or designing new ones. The idea is to describe mechanisms at a higher level of abstraction, focusing not on their implementation but on the *properties* they satisfy. In mathematics, abstraction extracts the underlying structures of a concept; in computer science, it generalises from concrete details to reusable specifications. Principle-based analysis applies this idea to reasoning.

This approach has a long tradition across domains. In voting theory, Arrow's axioms [8] specify criteria for voting systems and yield impossibility results. In belief revision, AGM postulates ensure the rationality of revision operations [2]. In nonmonotonic logic, Gabbay discussed the central requirements reflexivity, cut, and cautious monotony [44]. In formal argumentation, the principle-based view has been applied at the different stages of the Inference-as-Argumentation RAD in Figure 2. For arrow (1), the Kraus–Lehmann–Magidor (KLM) principles [57] axiomatise properties a nonmonotonic consequence relation ought to satisfy. For arrows (2)–(4), axiomatic analyses compare *attack assignments* among arguments [36, 38, 59]. For the whole RAD, *rationality postulates* guarantee that the overall conclusions satisfy desirable properties such as direct/indirect consistency and closure [24, 25, 23].

For abstract argumentation, the semantics has been classified via principles [11], with further extensions in [65]. The same methodology carries over to extended argumentation frameworks, yielding principle-based analyses for ranking-based semantics [4], gradual semantics [12, 18], multi-agent argumentation [69], and bipolar argumentation [68], etc.

Principle-based approach supports two main aims. First, it can be used to *define new input–output behaviours* of reasoning by selecting the principles one wishes to enforce as desiderata. Second, it can be used to *compare existing input–output behaviours* by identifying which principles they satisfy or fail to satisfy. Beyond comparison, principle-based analysis can yield *characterisation theorems*, where a given set of principles uniquely determines a function, and *impossibility results*, which show that no function can satisfy certain sets of principles simultaneously.

In the age of agentic AI, the same style of analysis must also apply to *subsymbolic* components. Even if a component is opaque internally, we can still analyse its input–output behaviour against principles. This connects naturally with ongoing work on verification of machine learning models [55]: for example, proving that a classifier respects monotonicity, fairness, or robustness constraints. In this way, the principle-based approach provides a unifying layer for both symbolic and subsymbolic components. It extends the toolbox idea: principles allow us not only to combine mechanisms but also to constrain and govern them, ensuring that even black-box modules contribute to explainable and accountable agentic systems.

5 Related work

We introduced reasoning alignment diagrams (RADs) as an architectural pattern for aligning different forms of reasoning. Our main technical contribution is the introduction of the revision-as-argumentation RAD and a proof of alignment for the specific case of base revision and base argumentation. The connection between argumentation and belief revision has been widely recognised in the literature, grounded in the view that argumentation is an inherently dynamic process [9, 40, 58]. Both frameworks aim to model rational change in light of new information, and argumentation has been proposed as a natural mechanism for capturing belief dynamics and resolving inconsistencies. Under this perspective, belief addition corresponds to introducing new arguments, contraction may involve removing arguments or introducing counterarguments, and the evolving structure of the argumentation framework reflects the agent's shifting epistemic state.

This conceptual link has been explored from multiple directions. A substantial

body of work investigates change in abstract argumentation frameworks, examining how adding or removing arguments and attacks affects extension sets under various semantics [14, 13, 26, 29, 33, 34, 64]. Other lines of research model change in argumentation systems explicitly as a form of belief revision, drawing analogies with AGM-style postulates and operators [21, 15, 54]. Structured argumentation has also been studied in this context, with attention to how changes to the underlying knowledge base or rule set relate to changes in the resulting argument graph [16, 61, 62]. Despite these efforts, representation results for belief revision as a form of formal argumentation are lacking. In Section 3 we provided such a representation result establishing a direct correspondence between full-meet belief base revision and base argumentation. To the best of our knowledge, this is the first work to show that this form of belief revision can be naturally represented within a structured argumentation framework.

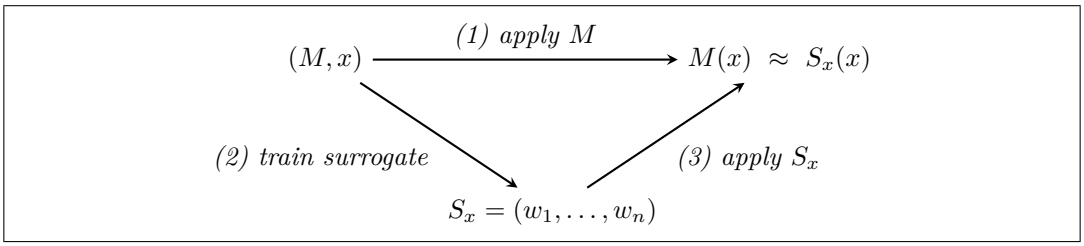

Figure 7: The *Interpretable Surrogate* RAD: prediction $M(x)$ of an opaque model M (arrow 1) is explained by training a linear model S_x (arrows 2, 3). The coefficients $w_1, \ldots, w_n$ of S_x represent feature importance values for the prediction $M(x)$.

In this paper we used RADs to represent different forms of symbolic reasoning, where one route provides an argumentation-based explanation for another. Yet the RAD idea is more general and also applies to explanation in subsymbolic approaches. A good example is the *surrogate-model* approach popularised by LIME in the field of Explainable AI [63] (see Figure 7). Given an opaque model M (e.g., a neural network) and an input x, the aim is to provide an explanation for the prediction $M(x)$. One route simply computes $M(x)$ without providing any explanation. The alternative route proceeds in two steps: first, an interpretable surrogate S_x is trained to be *locally faithful* to M (meaning that $M(x) \approx S_x(x)$ for inputs 'close to' x); second, the prediction $S_x(x)$ is obtained, where local faithfulness ensures (approximate) alignment of the RAD. The point of the alternative route is not to provide an alternative way to obtain the prediction, but rather to obtain explanations for the prediction made by the source model. This is typically done by using a linear classifier for the interpretable surrogate, and using the regression

coefficients $(w_1, \ldots, w_n)$ as feature importance values. This example shows that the RAD concept can capture explanation even in purely subsymbolic settings.

6 Summary and Conclusion

Our aim was to advance Dov Gabbay's vision of *logic as a toolbox*, of *disciplined combinations*, and of *argumentation as the host formalism* by developing a unifying pattern for aligning different forms of reasoning based on *Reasoning Alignment Diagrams* (RADs). RADs are commutative reasoning representations in which distinct routes from input to output align, enabling the separation of a *normative path* that specifies what is sanctioned, and an *argumentation path* that explains why. The pattern admits exact, partial, and approximate alignment. Routes consisting of multiple steps make explicit the translation from knowledge bases to arguments and attacks, the choice of Dung-style semantics, and the extraction of conclusions from extensions.

We showed that existing work grounded in Dung's model fits the *Inference-as-Argumentation* RAD, which explains defeasible logics via the structured argumentation route (construct an AF, apply semantics, extract conclusions). *Argumentation-as-Discussion* adds a further explanatory layer, by recasting acceptance under a semantics in terms of two-player discussion games with proponent/opponent strategies. Together, these perspectives illustrate how argumentation provides both an abstraction across heterogeneous formalisms and an explanation engine for their outcomes.

Our new contribution is the *Revision-as-Argumentation* RAD that extends this role beyond inference to belief change. We considered full-meet belief *base* revision of a consistent K by φ and showed exact alignment with base argumentation. The construction builds a base-AF over $K \cup \varphi$ and introduces a *restricted-attack* relation that disables attacks originating from arguments contained in a chosen remainder set. We proved that the revised base coincides with the intersection of the premises that appear in all stable extensions of the modified framework. Intuitively, contraction by $\neg\varphi$ is realised by restricting attacks, while expansion by φ is captured by including φ in the base. This approach provides, to our knowledge, the first exact representation of full-meet base revision within base argumentation.

Methodologically, the RAD concept yields three pay-offs. First, it couples *specification and explanation*: the commutation amounts to a correctness guarantee of the explanation route. Second, it supports *modularity*, since Dung-style semantics can be reused across source formalisms. Third, it ensures *transparency*: attack/defence structures make the survival and removal of premises during revision explicit. These

benefits instantiate Gabbay's principles: mechanisms rather than monoliths, disciplined combinations, and argumentation as the operating system for heterogeneous reasoning.

We also introduced the idea of composing RADs. At the level of a single agent, combining argumentation-as-inference with revision-as-argumentation clarifies choice points (whether to assert or to listen; what to assert; and whether to revise, ignore, question, or challenge) and allows qualitative decision models to govern these communicative acts. At the dialogue level, composing the *assert* RAD (generation of claims from a base) with the *listen* RAD (revision of the receiver's base) yields explainable and auditable multi-agent interaction over time. We furthermore showed that the RAD concept applies beyond symbolic reasoning, for example by modelling subsymbolic surrogate-based explanation (e.g., LIME) as a surrogate-model RAD. We also envisioned that RADs can be extended to AI reasoning in the context of agentic AI, where symbolic and subsymbolic reasoning can be combined, and a principle-based methodology for specifying and comparing Multi-RAD architectures at their input–output behaviors.

Concerning the notion of revision-as-argumentation, two strands of future work follow. Firstly, Theorem 3.12 can be generalized by considering other kinds of revision operators, including iterated revision, prioritized revision, and external revision (see, e.g., [41] for a discussion of these variants). Furthermore, the problem of revising rules will be interesting to extend to other forms of structured argumentation, such as assumption-based argumentation [37] and ASPIC+ [60]. Moreover, extended forms of argumentation can be explored like ranking semantics [4], bipolar argumentation [27], and multiagent argumentation [7, 19]. Secondly, Theorem 3.12 assumes an initially consistent base. In practice, when an agent with an inconsistent knowledge base receives new information and revises its beliefs, some conflicts may be resolved, but it is unlikely that all inconsistencies will disappear. Consequently, we should consider other revision operators that may have inconsistent knowledge bases as a result [53]. Moreover, the revision of inconsistent knowledge bases can be driven by research on inconsistency measures [6]. The success measure of traditional belief revision, which says that the revised knowledge base must be consistent, can be replaced by a condition that the inconsistency measure of the knowledge bases does not increase. Other success conditions can be given, stating conditions under which the inconsistency measure must decrease.

Acknowledgments

L. van der Torre and L. Yu acknowledge the financial support of the Luxembourg National Research Fund (FNR). L. van der Torre is supported through the projects *The Epistemology of AI Systems (EAI)* (C22/SC/17111440), *DJ4ME – A DJ for Machine Ethics: the Dialogue Jiminy* (O24/18989918/DJ4ME), and *Logical Methods for Deontic Explanations (LoDEx)* (INTER/DFG/23/17415164/LoDEx). L. van der Torre and L. Yu are supported through the project *Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN)* (C24/IS/19003061/ SERAFIN). L. Yu is additionally supported by the University of Luxembourg's *Marie Speyer Excellence Grant* (2024) for the project *Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).*

References

[1] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. Agentic AI: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 2025.

[2] Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2):510–530, 1985.

[3] Carlos E Alchourrón and David Makinson. On the logic of theory change: Contraction functions and their associated revision functions. *Theoria*, 48(1):14–37, 1982.

[4] Leila Amgoud and Jonathan Ben-Naim. Ranking-based semantics for argumentation frameworks. In *International Conference on Scalable Uncertainty Management*, pages 134–147. Springer, 2013.

[5] Charles Antaki and Ivan Leudar. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2):181–194, 1992.

[6] Ofer Arieli, Kees van Berkel, Badran Raddaoui, and Christian Straßer. Deontic reasoning based on inconsistency measures. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 21, pages 71–81, 2024.

[7] Ryuta Arisaka, Ken Satoh, and Leendert van der Torre. Anything you say may be used against you in a court of law: Abstract agent argumentation (Triple-A). In *International Workshop on AI Approaches to the Complexity of Legal Systems*, pages 427–442. Springer, 2015.

[8] Kenneth J. Arrow. *Social Choice and Individual Values*, volume 12 of *Cowles Foundation Monographs*. Yale University Press, New Haven, CT, 1951.

[9] Pietro Baroni, Eduardo Fermé, Massimiliano Giacomin, and Guillermo Ricardo Simari. Belief revision and computational argumentation: A critical comparison. *J. Log. Lang. Inf.*, 31(4):555–589, 2022.

[10] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors. *Handbook of Formal Argumentation*, volume 1. College Publications, 2018.

[11] Pietro Baroni and Massimiliano Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10-15):675–700, 2007.

[12] Pietro Baroni, Antonio Rago, and Francesca Toni. How many properties do we need for gradual argumentation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[13] Ringo Baumann. What does it take to enforce an argument? minimal change in abstract argumentation. In Luc De Raedt, Christian Bessiere, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, editors, *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 127–132. IOS Press, 2012.

[14] Ringo Baumann and Gerhard Brewka. Expanding argumentation frameworks: Enforcing and monotonicity results. In Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Guillermo Ricardo Simari, editors, *Computational Models of Argument: Proceedings of COMMA 2010, Desenzano del Garda, Italy, September 8-10, 2010*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 75–86. IOS Press, 2010.

[15] Ringo Baumann and Gerhard Brewka. AGM meets abstract argumentation: Expansion and revision for dung frameworks. In Qiang Yang and Michael J. Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2734–2740. AAAI Press, 2015.

[16] Matti Berthold, Anna Rapberger, and Markus Ulbricht. Forgetting aspects in assumption-based argumentation. In Pierre Marquis, Tran Cao Son, and Gabriele Kern-Isberner, editors, *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023*, pages 86–96, 2023.

[17] Philippe Besnard and Anthony Hunter. A review of argumentation based on deductive arguments. *Handbook of Formal Argumentation*, 1(437-484):111, 2018.

[18] Vivien Beuselinck, Jérôme Delobelle, and Srdjan Vesic. A principle-based account of self-attacking arguments in gradual semantics. *Journal of Logic and Computation*, 33(2):230–256, 2023.

[19] Elizabeth Black, Nicolas Maudet, and Simon Parsons. Argumentation-based dialogue. *Handbook of Formal Argumentation, Volume 2*, 2021.

[20] Richard Booth, Souhila Kaci, Tjitze Rienstra, and Leendert van Der Torre. A logical theory about dynamics in abstract argumentation. In *Scalable Uncertainty Management: 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings 7*, pages 148–161. Springer, 2013.

[21] Richard Booth, Souhila Kaci, Tjitze Rienstra, and Leendert W. N. van der Torre. A

logical theory about dynamics in abstract argumentation. In Weiru Liu, V. S. Subrahmanian, and Jef Wijsen, editors, *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*, volume 8078 of *Lecture Notes in Computer Science*, pages 148–161. Springer, 2013.

[22] Martin Caminada. Argumentation semantics as formal discussion. *Handbook of Formal Argumentation*, 1, 2017.

[23] Martin Caminada. Rationality postulates: Applying argumentation theory for non-monotonic reasoning. *Handbook of Formal Argumentation, Volume 1*, pages 771–796, 2018.

[24] Martin Caminada and Leila Amgoud. An axiomatic account of formal argumentation. In *AAAI*, volume 6, pages 608–613, 2005.

[25] Martin Caminada and Jonathan Ben-Naim. *Postulates for paraconsistent reasoning and fault tolerant logic programming.* PhD thesis, Department of Information and Computing Sciences, Utrecht University, 2007.

[26] Claudette Cayrol, Florence Dupin de Saint-Cyr, and Marie-Christine Lagasquie-Schiex. Change in abstract argumentation frameworks: Adding an argument. *CoRR*, abs/1401.3838, 2014.

[27] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. Bipolar abstract argumentation systems. In *Argumentation in Artificial Intelligence*, pages 65–84. Springer, 2009.

[28] Jinsheng Chen, Beishui Liao, and Leendert van der Torre. Bisimulation between base argumentation and premise-conclusion argumentation. *Artificial Intelligence*, 336:104203, 2024.

[29] Sylvie Coste-Marquis, Sébastien Konieczny, Jean-Guy Mailly, and Pierre Marquis. On the revision of argumentation systems: Minimal change of arguments statuses. In Chitta Baral, Giuseppe De Giacomo, and Thomas Eiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*. AAAI Press, 2014.

[30] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative XAI: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4392–4399. ijcai.org, 2021.

[31] Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. *Artificial intelligence*, 89(1-2):1–29, 1997.

[32] Mehdi Dastani, Joris Hulstijn, and Leendert Van der Torre. How to decide what to do? *European Journal of Operational Research*, 160(3):762–784, 2005.

[33] Sylvie Doutre, Andreas Herzig, and Laurent Perrussel. Abstract argumentation in dynamic logic: Representation, reasoning and change. In Beishui Liao, Thomas Ågotnes, and Yì N. Wáng, editors, *Dynamics, Uncertainty and Reasoning, The Second Chinese Conference on Logic and Argumentation, CLAR 2018, Hangzhou, China, 16-17 June 2018*, pages 153–185. Springer, 2018.

[34] Sylvie Doutre and Jean-Guy Mailly. Constraints and changes: A survey of abstract

argumentation dynamics. *Argument Comput.*, 9(3):223–248, 2018.

[35] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[36] Phan Minh Dung. An axiomatic analysis of structured argumentation with priorities. *Artif. Intell.*, 231:107–150, 2016.

[37] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. Assumption-based argumentation. In *Argumentation in artificial intelligence*, pages 199–218. Springer, 2009.

[38] Phan Minh Dung and Phan Minh Thang. Fundamental properties of attack relations in structured argumentation with priorities. *Artificial Intelligence*, 255:1–42, 2018.

[39] Artur S d'Avila Garcez, Luis C Lamb, and Dov M Gabbay. *Neural-symbolic cognitive reasoning*. Springer, 2009.

[40] Marcelo Alejandro Falappa, Gabriele Kern-Isberner, and Guillermo Ricardo Simari. Belief revision and argumentation theory. In Guillermo Ricardo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 341–360. Springer, 2009.

[41] Eduardo Fermé and Sven Ove Hansson. *Belief change: introduction and overview*. Springer, 2018.

[42] FIPA. Communicative act library specification. *http://www. fipa. org/specs/fipa00037*, 2002.

[43] Andrew Fuchs, Andrea Passarella, and Marco Conti. A cognitive framework for delegation between error-prone ai and human agents. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 317–322. IEEE, 2022.

[44] Dov M. Gabbay. Theoretical foundations for non-monotonic reasoning in expert systems. In Krzysztof R. Apt, editor, *Logics and Models of Concurrent Systems*, pages 439–457. Springer-Verlag, New York, 1985.

[45] Dov M. Gabbay. Fibred semantics and the weaving of logics. part 1: Modal and intuitionistic logics. *The Journal of Symbolic Logic*, 61(4):1057–1120, 1996.

[46] Dov M Gabbay. *Labelled deductive systems*. Oxford university press, 1996.

[47] Dov M Gabbay. An overview of fibred semantics and the combination of logics. In *Frontiers of Combining Systems: First International Workshop, Munich, March 1996*, pages 1–55. Springer, 1996.

[48] Dov M Gabbay. *Fibring logics*, volume 38. Clarendon Press, 1998.

[49] Dov M Gabbay. Fibring argumentation frames. *Studia Logica*, 93(2):231, 2009.

[50] Dov M Gabbay and Sérgio Marcelino. Modal logics of reactive frames. *Studia Logica*, 93(2):405, 2009.

[51] Dov M Gabbay and Lydia Rivlin. Heal2100: human effective argumentation and logic for the 21st century. the next step in the evolution of logic. *IFCoLog Journal of Logics and Their Applications*, 2017.

[52] Sven Ove Hansson. *A textbook of belief dynamics - theory change and database updating*, volume 11 of *Applied logic series*. Kluwer, 1999.

[53] Sven Ove Hansson and Renata Wassermann. Local change. *Studia Logica*, 70:49–76, 2002.

[54] Adrian Haret, Johannes Peter Wallner, and Stefan Woltran. Two sides of the same coin: Belief revision and enforcing arguments. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 1854–1860. ijcai.org, 2018.

[55] Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pages 749–758. PMLR, 2020.

[56] Hirofumi Katsuno and Alberto O Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1991.

[57] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial intelligence*, 44(1-2):167–207, 1990.

[58] Fabio Paglieri and Cristiano Castelfranchi. Revising beliefs through arguments: Bridging the gap between argumentation and belief revision in MAS. In Iyad Rahwan, Pavlos Moraitis, and Chris Reed, editors, *Argumentation in Multi-Agent Systems, First International Workshop, ArgMAS 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers*, volume 3366 of *Lecture Notes in Computer Science*, pages 78–94. Springer, 2004.

[59] Pere Pardo, Liuwen Yu, Chen Chen, and Leendert van der Torre. Weakest link, prioritized default logic and principles in argumentation. *Journal of Logic and Computation*, 35(4):exaf007, 2025.

[60] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.

[61] Henry Prakken. Relating abstract and structured accounts of argumentation dynamics: the case of expansions. In Pierre Marquis, Tran Cao Son, and Gabriele Kern-Isberner, editors, *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023*, pages 562–571, 2023.

[62] Anna Rapberger and Markus Ulbricht. On dynamics in structured argumentation formalisms. *J. Artif. Intell. Res.*, 77:563–643, 2023.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[64] Tjitze Rienstra, Chiaki Sakama, Leendert van der Torre, and Beishui Liao. A principle-based robustness analysis of admissibility-based argumentation semantics. *Argument Comput.*, 11(3):305–339, 2020.

[65] Leendert van der Torre and Srdjan Vesic. The principle-based approach to abstract argumentation semantics. In Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors, *Handbook of Formal Argumentation, Volume 1*, pages 797–838. College Publications, 2018.

[66] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.*, 36:e5, 2021.

[67] Douglas Walton and Erik CW Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, 1995.

[68] Liuwen Yu, Caren Al Anaissy, Srdjan Vesic, Xu Li, and Leendert van der Torre. A principle-based analysis of bipolar argumentation semantics. In *European Conference on Logics in Artificial Intelligence*, pages 209–224. Springer, 2023.

[69] Liuwen Yu, Dongheng Chen, Lisha Qiao, Yiqi Shen, and Leendert van der Torre. A Principle-based Analysis of Abstract Agent Argumentation Semantics. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning*, pages 629–639, 11 2021.

[70] Liuwen Yu and Leendert van der Torre. The A-BDI metamodel for human-level AI: Argumentation as balancing, dialogue and inference. In *International Conference on Logic and Argumentation (CLAR 2025)*, Cham, 2025. Springer Nature Switzerland. To appear.

[71] Liuwen Yu, Leendert Van der Torre, and Réka Markovich. Thirteen challenges in formal and computational argumentation. *Handbook of Formal Argumentation*, 3:931–1012, 2024.

 Received September 2025