DiSCo–RAD: Reasoning Alignment for Judicial Discretion

Liuwen Yu[1], Leendert van der Torre[2,3], Réka Markovich[2], Beishui Liao[3], Chenyang Cai[3]

[1] *Luxembourg Institute of Science and Technology, Luxembourg*
[1] *University of Luxembourg, Luxembourg*
[2] *Zhejiang University, China*

Abstract

We analyse how judicial majorities can emerge despite internal disagreement through a model called DiSCo–RAD, developed on a real case from the Hungarian Constitutional Court. In this case, the judges agreed on the outcome but expressed distinct views through *parallel reasonings* and *dissenting opinions*, illustrating the discretionary nature of constitutional adjudication. In our model, the term *coalition* is used metaphorically to describe a convergent group of judges—a *majority configuration* that reflects agreement in outcome. DiSCo–RAD keeps two routes in view and checks whether they converge on the same majority constellation. On the *normative route*, judges are represented as a small social network built from pairwise assessments of cooperation. From these relations—friend, enemy, or mixed—we derive groups of compatible reasoning by a "friend-first with conflict tolerance" rule that allows limited disagreement when allies can compensate for it. On the *argumentation route*, possible collaborations are represented as nodes in a structured argumentation framework, with attacks expressing incompatible choices and supports expressing tolerated disagreement. When the reasoning patterns obtained from both routes coincide, the diagram explains how the majority holds together: differences in reasoning remain within the bounds of mutual support rather than full consensus.

Keywords: Artificial Intelligence, Knowledge representation and reasoning, Judicial discretion, Formal argumentation, Reasoning alignment, Conflict tolerance

1 Introduction

Constitutional courts are a special case of *discretionary judicial reasoning*. Their task is not merely to apply statutes, but to ensure that legislation and judicial decisions conform to the constitution and to the principles that sustain it—such as equality, proportionality, and the protection of fundamental rights. Because these principles may conflict, the reasoning of constitutional judges necessarily involves *discretion*: they interpret open-textured norms, weigh competing values, and justify why one interpretation better preserves constitutional coherence. Hence, *agreement on the outcome does not entail agreement on the*

reasoning. Judges may attach separate opinions—*dissenting opinions* expressing disagreement with the decision, and *parallel reasonings* expressing agreement with the result but not with the majority's rationale [1]. These practices make constitutional courts a privileged field for studying how collective legal reasoning accommodates disagreement. We use the word "*coalition*" metaphorically to denote a convergent group of judges—a majority configuration based on compatible reasoning, not a political alliance.

Formal argumentation provides a natural lingua franca for such explanations. Rather than introduce yet another universal formalism, we adopt the "logic as a toolbox" view [23,34]: choose the mechanisms that fit the task, combine them transparently, and make their interaction auditable. In this paper, the toolbox contains three layers:

(1) a social layer that derives *friend*, *enemy*, and *frenemy* relations from pairwise collaboration scores between judges, reflecting how similar or divergent their reasoning is;

(2) an argumentation layer that encodes each judge's possible collaborations as *arguments*, where *attacks* represent incompatible choices (a judge cannot coherently adopt both), and *supports* represent tolerable disagreement (allies whose reasons partly cover a conflict);

(3) a coalition layer that extracts majorities (*Maj*) and checks individual rationality (*IR*, meaning no mutually harmful pairs).

We introduce **DiSCo–RAD**, a *Reasoning Alignment Diagram*, which keeps two reasoning routes side by side and checks that they match.

(1) Normative route. From collaboration values $\pi(\alpha, \beta)$ we classify each pair of judges as friends, enemies, or frenemies, then grow coalitions by a *friend-first + tolerance* closure. Tolerance is governed by a simple margin k: an ally can *cover* a local disagreement if $\pi(\alpha, \beta) + \pi(\alpha, \gamma) \geq k$. The result is filtered by *Maj* (majority threshold) and *IR* (no enemies inside). In legal terms, this route models how a majority can hold when minor disagreements are compensated by sufficient mutual understanding among judges.

(2) Argumentation route. We construct a *contrastive bipolar argumentation framework (BAF)* [2] whose arguments are directed collaborations (α, β). The *attack* relation is *local*: for a fixed judge α, (α, β) *attacks* (α, γ) whenever $\pi(\alpha, \gamma) < 0$, meaning that cooperation with γ would harm α's reasoning consistency. The *support* relation encodes the tolerance rule: (β, α) *supports* (α, γ) when α and β are friends, $\pi(\alpha, \gamma) < 0$ and $\pi(\alpha, \beta) + \pi(\alpha, \gamma) \geq k$. Acceptability is evaluated using the *β-semantics* for bipolar frameworks with *deductive support*, where accepted supporters propagate acceptance to what they support, and defeat requires neutralising all supporters. When support is ignored, this collapses to Dung's classic semantics [22]. In judicial terms, this route represents how compatible interpretations reinforce one another while still per-

[1] Section 66 of Act CLI of 2011 on the Constitutional Court of Hungary

mitting disagreement within defensible limits.

We then *extract* coalitions from accepted collaborations by checking *mutual intent* (both (α, β) and (β, α) appear) and forming tolerance-aware cliques. A final *IR-pruning* step removes any residual enemies without losing the tolerance witnesses already present. The diagram itself provides the explanation: when both routes produce the same group of judges, the diagram *commutes*, meaning that the computational reasoning aligns with the normative account of discretion.

We validate DiSCo–RAD on a constitutional-court case. The numerical values behind π are illustrative—derived from judicial opinions rather than measured preferences—yet they reproduce the observed majority and explain why a partially disagreeing judge remains in it: supportive allies provide enough margin to cover the disagreement.

The remainder of this paper is structured as follows. Section 2 presents the case and inputs. Section 3 gives the methodology and the diagram. Section 4 formalises the social layer, the contrastive BAF with local attacks and deductive supports, and the coalition extractor. Section 5 works through the case. Section 6 discusses related work, and Section 7 concludes.

2 Discretionary Case Study

Constitutional courts are collegial bodies that review whether statutes—and, in some systems, certain judicial decisions—comply with the constitution. Cases typically reach them through a constitutional complaint, petition, or judicial referral; they are not ordinary appellate courts in the sense of re-examining facts or evidence. Once the case file is prepared, the bench (either a panel or the full court) deliberates *behind closed doors*: judges exchange drafts and reasons, discuss the constitutional issues, and finally vote on the operative part of the decision—such as to annul, uphold, or dismiss the challenged provision or complaint. These deliberations are confidential. What becomes public is the final decision (the operative part and the reasoning of the majority) together with any *separate opinions*.

Separate opinions come in two main forms. A *dissenting opinion* is written by a judge who disagrees with the decision's outcome or its reasoning and wishes to record the grounds of that disagreement. A *parallel reasoning* is written by a judge who agrees with the decision's outcome but disagrees on the reasoning. Both types of opinions serve transparency and doctrinal development by revealing the diversity of constitutional interpretation within the court.

Because internal exchanges are confidential, our model is *normative and explanatory*, not descriptive. We reconstruct a defensible pathway that *could* account for the published outcome and the pattern of separate opinions. Accordingly, the collaboration values in Table 1 and the induced win–lose signs in Table 2 are *illustrative encodings* derived from the published reasons; they are

not measured preferences or survey data and should be read only as for explanation. The numerical entries of π and the sign profiles are therefore explicitly *for illustration* and to make the explanation auditable.

Our approach sits within a line of work that treats judicial discretion as a normative reasoning problem with explicit freedom and obligations. In particular, Dik & Markovich formalise [20,19,21] a *duty of care* as the boundary of discretionary freedom in child custody cases: judges retain freedom to choose between the parents, but are constrained by obligations to determine relevant facts, to weigh them, and to reason consistently. We draw on that perspective in positioning the coalition-level postulates used later (IR, Tol(k), Maj) as boundaries on tolerable internal disagreement.

2.1 The Hungarian Constitutional Court case

To illustrate our approach, we consider Decision 3023/2016 (II. 23.) of the Hungarian Constitutional Court, which concerned maternity benefits and the interpretation of the Social Insurance Act (LXXXIII/1997). The petitioner, a mother who moved from one full-time to two part-time contracts before childbirth, was denied maternity benefit for one of the jobs. According to § 43(2) of the Act, each concurrent insurance relationship must independently satisfy a 365-day contribution period. The mother argued that this rule unfairly penalised her because she had paid social contributions for both jobs and should receive equal coverage.

The case therefore turned on how to weigh *formal equality under the law* against *substantive fairness of outcomes*. The constitutional judges were deeply divided. Fourteen judges (one absent) issued seven dissenting, four concurring, and only three purely concurring opinions. The majority upheld the statute, emphasising predictability and legal certainty, while a strong minority considered the rule discriminatory against women in non-standard employment. These divisions make the case a natural laboratory for studying *conflict-tolerant coalitions*: the court reached a majority decision even though complete agreement was impossible.

2.2 Agents and their simplified arguments

We model the judges as five representative *agents* ($Ag = \{\alpha, \beta, \gamma, \delta, \varepsilon\}$) that capture the main interpretive positions observed in the case:

- α – a legal formalist who stresses constitutionality and uniform application of statutes.
- β – another formalist who focuses on textual fidelity and legal certainty, warning against vague moral criteria.
- γ – a moderate intermediary: accepts the law's constitutionality but finds the outcome unjust, proposing a more value-sensitive reading.
- δ – a reform-minded judge arguing that the rule produces discrimination and should be struck down.
- ε – a fairness-oriented judge supporting δ, claiming that formal categories ignore substantive equality.

Concrete arguments a_1–a_6. Let $Ar = \{a_1, \ldots, a_6\}$ with:

- a_1 (by α): "The rule (§43(2) Ebtv.) is constitutional and applies equally; no discrimination."
- a_2 (by α): "Courts should avoid undefined moral terms (just/unjust) in constitutional review."
- a_3 (by β): "Stick to positive law and legal certainty; open moral notions undermine predictability."
- a_4 (by γ): "The rule is constitutional, but the outcome is unjust; adopt a value-sensitive interpretation (Art. 28)."
- a_5 (by δ): "The rule is discriminatory (Art. XV) against two-part-time contributors and should be struck down."
- a_6 (by ε): "Equal contributions but unequal benefit; formal classification ignores substantive equality."

Each triple (a_i, ω, ϵ) will later be used as a *reason* for agent ω to collaborate with ϵ when constructing the weighted base $\langle R, w \rangle$.

2.3 From arguments to collaboration values

Every ordered pair of judges (x, y) receives a *relative collaboration value* $\pi(x, y)$ measuring the net benefit that x expects from working with y. Positive values mean that x sees y as a helpful ally; negative values mean that x experiences y as an obstacle. The matrix in Table 1 summarises these relations.

$\pi(x,y)$	$y=\alpha$	$y=\beta$	$y=\gamma$	$y=\delta$	$y=\varepsilon$
$x=\alpha$	*	3	−2	−6	−8
$x=\beta$	2	*	−1	−3	−3
$x=\gamma$	2	1	*	−3	−3
$x=\delta$	−4	−3	−3	*	3
$x=\varepsilon$	−4	−3	−3	2	*

Table 1
Relative collaboration values $\pi(x, y)$ derived from the arguments.

In Table 1, rows indicate the agent that evaluates others; columns indicate the potential partner. For example, $\pi(\alpha, \beta) = 3$ shows that α perceives collaboration with β as strongly positive, while $\pi(\alpha, \gamma) = -2$ signals mild disagreement. Symmetry is not required: $\pi(\beta, \alpha) = 2$ is positive but not equal to $\pi(\alpha, \beta) = 3$, meaning that β values α slightly less. The first three rows correspond to the formalist camp: they cooperate internally (α and β) but have mild friction with γ. Rows for δ and ε show high mutual support ($\pi(\delta, \varepsilon) = 3$, $\pi(\varepsilon, \delta) = 2$) and strong opposition to the formalists (negative values across the first three columns).

Overall, two cohesive clusters appear: (α, β, γ) with mostly positive or small negative links, and (δ, ε) with mutual positive links but hostility to the first group. The contrastive structure of the network is already visible here.

2.4 Simplifying with win–lose signs

The sign transformation $tr(x)$ simplifies π into two values per pair: "+" if cooperation is favourable, "−" otherwise. We call the resulting tuples (v_1, v_2) the *win–lose profile* between two agents. The outcome is shown in Table 2.

$win{-}lose(x,y)$	$y=\alpha$	$y=\beta$	$y=\gamma$	$y=\delta$	$y=\varepsilon$
$x=\alpha$	*	$(+,+)$	$(-,+)$	$(-,-)$	$(-,-)$
$x=\beta$	$(+,+)$	*	$(-,+)$	$(-,-)$	$(-,-)$
$x=\gamma$	$(+,-)$	$(+,-)$	*	$(-,-)$	$(-,-)$
$x=\delta$	$(-,-)$	$(-,-)$	$(-,-)$	*	$(+,+)$
$x=\varepsilon$	$(-,-)$	$(-,-)$	$(-,-)$	$(+,+)$	*

Table 2
Win–lose outcomes after applying the sign function $tr(x)$.

Each cell (x, y) in Table 2 records the pair of signs: the first symbol represents how x evaluates y, the second how y evaluates x. The pattern $(+,+)$ indicates a *friend relation*, $(-,-)$ an *enemy relation*, and mixed signs $(+,-)$ or $(-,+)$ a *frenemy relation*. In the table, α and β form a strong friendship ($(+,+)$ in both directions). They both have asymmetric frenemy relations with γ, and mutual enmity with δ and ε. Conversely, δ and ε form a second friendship pair, isolated from the rest by $(-,-)$ relations.

These patterns summarise the intuition of the court's deliberation: a cooperative bloc of formalists (α, β, γ) opposed to a smaller but cohesive fairness bloc (δ, ε). The slightly negative values between γ and the other formalists capture the reality that Judge γ partly agreed on the outcome but disagreed on the moral reasoning—a form of *tolerated dissent*.

The discretion case demonstrates the need for a model that can explain coalitions that are *stable but not unanimous*. Traditional game-theoretic or logic-based models require complete alignment of utilities or rules, whereas courts and other multi-agent institutions rely on *partial alignment*: members may disagree on values yet still form a workable majority. The numerical values in Tables 1–2 encapsulate this phenomenon in a minimal form. They will serve as input to the weighted social network and contrastive argumentation framework analysed in Sections 3–4.

3 Methodology: A-BDI and DiSCo–RAD

We use *formal and computational argumentation* in the sense defined in the third volume of *Handbook of Formal Argumentation* [39]. In that usage, *formal argumentation* is the representation, management and (at times) resolution of conflict; *computational argumentation* studies and implements those processes with computational methods (algorithms, complexity, data structures) and their integration with other technologies.

Dung's abstract argumentation theory [22] is marked as an attack-defense paradigm shift: acceptance depends on how arguments *attack* and *defend* one another, abstracting from internal structure. In this view, argumentation is a

graph-based abstraction of nonmonotonic inference. This perspective has become a central bridge across AI subfields, because the same abstract machinery can be specialised, extended, compared, and aligned with other reasoning formalisms.

Within this context, argumentation has been developed along three complementary *conceptualizations* that we will later align with the A–BDI metamodel: argumentation as *balancing*, as *dialogue*, and as *inference*. Over the past two decades this has yielded a family of argumentation representations:

Value-based and weighted models enrich abstract argumentation with priorities or numeric strengths to articulate trade-offs among competing reasons [13,35].

Bipolar and coalitional models distinguish support from attack and capture cooperation/opposition among agents (early contributions from the Toulouse school and collaborators) [3,17,14].

Conflict tolerant semantics relax conflict-freeness to model reasoning under inconsistency or disagreement [7,8].

Multi-agent and dialogical approaches view argumentation as interaction (debate, persuasion, negotiation), often linked to social choice and deliberation [10,9].

The Handbook of Formal Argumentation volumes survey this landscape in depth [25,13,39]. In this paper, rather than treating diversity as a problem, we adopt the *logic-as-toolbox* stance: combine components (semantics, weighting, aggregation, dialogue protocols) to suit the application, and verify the fit by *alignment*. In the remainder, we operationalise that stance with two devices. First, the *A–BDI* metamodel will organise the modelling layers. Second, *Reasoning Alignment Diagrams* (RADs) will make explicit how a normative route and an argumentation route commute—or where and why they diverge. Our case study on judicial discretion then instantiates this toolbox for coalition formation among judges and for explaining patterns of separate opinions.

3.1 The logic-as-toolbox perspective

In line with the tradition in the Knowledge Representation and Reasoning (KR) community, Gabbay [23] promotes the view of logic not as a single system but as a *toolbox of reasoning mechanisms*: each mechanism is a component that can be selected and combined for particular applications. This view motivates our modelling choices. Instead of introducing yet another new formalism, we assemble existing ideas into a transparent pipeline:

(1) A weighted reason base captures the balance of benefits and costs between agents;

(2) A social layer defines relations such as friendship, enmity, and support;

(3) An argumentation layer transforms these relations into a contrastive argumentation framework whose semantics determine accepted collaborations;

(4) A coalition layer extracts groups of agents that jointly satisfy rationality, tolerance, and majority postulates.

Each layer can be replaced or refined without affecting the others, mirror-

ing the modular design advocated in the Handbook of Formal Argumentation volumes [25,13,39].

3.2 The A–BDI metamodel

The A-BDI metamodel [38] organizes this toolbox into three complementary perspectives:

Balancing as identified by Gordon [25] involves weighing the pros and cons of an issue in order to reach a balanced decision or judgment. In such a system, pro and con arguments for alternative resolutions of the issues (options or positions) are put forward, evaluated, resolved, and balanced [25]. The formal methods used are multi-criteria decision theory and case-based reasoning, and they are applied in the law [26], ethics [35] and decision theory [18].

Dialogue Argumentation as dialogue conceptualizes argumentation as a form of interaction aimed at resolving conflicts of opinion [31]. It focuses on the exchange among multiple agents according to defined protocols. Dialogue highlights the distributed nature of information, the selective disclosure of arguments [9], and the agents' strategies for achieving collective or competing objectives [9].

Inference covers the logical semantics that determine which arguments, or collaborations, are ultimately accepted. Here we use Dung-style semantics [22] but adapted to a *contrastive* setting where attacks come from competing collaboration options of the same source.

In this paper, we interpret *dialogue* in a broader sense than conversational or turn-taking protocols. It refers to the *relational layer* where agents can be friends, enemies, or *frenemies* [36,27]. We define the network of mutual influences within which balancing operates, governing how agents' reasoning interacts for mutual or individual benefit, when disagreements can be shielded by coalitions, and how collective acceptance emerges from inter-agent relations; this interpretation echoes work on bipolar [2,17] and social argumentation [37,33,14].

The A-BDI view emphasises that these three layers are not sequential computations but complementary descriptions: each layer can be formalised and reasoned about in its own right. Together, they support modular design and explanation.

3.3 Reasoning Alignment Diagrams (RADs)

RAD visualises how different reasoning routes—for example, a normative specification and a computational procedure—can be aligned or "commuted". In a RAD, each route transforms the same input into an output. If both produce the same result, the diagram *commutes*: the system is sound and complete with respect to its specification. If not, the diagram reveals where approximation or information loss occurs.

RADs can also serve as a general framework for connecting the three conceptualizations of argumentation introduced earlier. They instantiate the *attack–defence paradigm shift*—Dung's abstract argumentation—by showing how

different forms of reasoning can be represented within a unified structure. For example, Dung showed that several forms of nonmonotonic inference can be represented within his theory of abstract argumentation. This correspondence can be depicted by the RAD in Figure 1 [39]. The diagram relates two approaches to deriving conclusions from a knowledge base. The first route (arrow 1) is the *canonical* inference route defined by the underlying defeasible logic. The second route (arrows 2–3–4) is the *argumentation route*: it begins with the translation of the knowledge base into a structured argumentation framework (arrow 2), applies semantics to determine extensions (arrow 3), and extracts the conclusions of the accepted arguments (arrow 4).

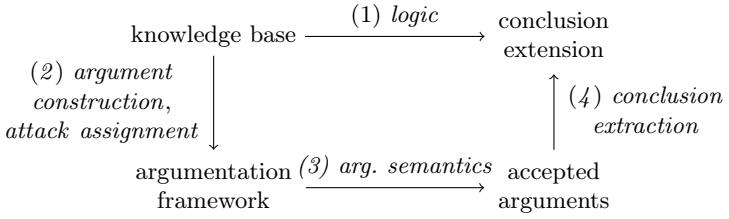

Fig. 1. The Inference-as-Argumentation RAD: the argumentation route (arrows 2–3–4) explains the inferences performed by the source formalism (arrow 1) [39].

RADs can also combine *argumentation as inference* and *argumentation as dialogue*. Consider the problem of determining the extensions of an argumentation framework under a given semantics (arrow 3 in Figure 1). This problem can be reformulated in terms of *two-player discussion games* [15], where a *proponent* and an *opponent* alternate in attacking or defending arguments. Starting from an initial claim by the proponent, the dialogue follows a fixed protocol; the existence of a winning strategy for the proponent corresponds to the argument being accepted. This correspondence is shown in the *Argumentation-as-Discussion* RAD in Figure 2. Here, arrow 1 represents the canonical evaluation of semantics, whereas arrows 2–3–4 describe its computation through a discussion game. Note that arrow 1 in Figure 2 corresponds to arrow 3 in Figure 1.

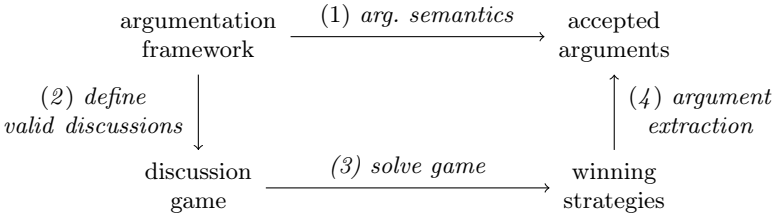

Fig. 2. The Argumentation-as-Discussion RAD: discussion games (arrows 2–3–4) explain argument acceptability (arrow 1) [1].

In this paper, we introduce the *DiSCo–RAD*. It aligns a *normative route*, defined by coalition postulates such as IR, Tol(k), and Maj, with a *compu-*

tational route expressed in a contrastive argumentation game. The diagram commutes when both routes yield the same coalition, showing that the computational mechanism faithfully realises the normative model of discretionary reasoning.

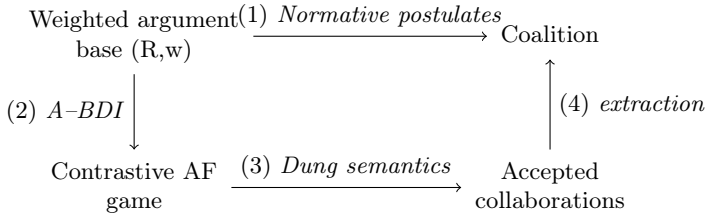

Fig. 3. DiSCo–RAD: alignment between the normative and argumentation routes.

3.4 Applying the methodology to discretion

We start from a legal case rather than from abstract definitions, because judicial discretion provides an exemplary domain where multiple reasoning models coexist and must be combined. Legal formalism relies on textual interpretation; value-based reasoning appeals to moral and social considerations; and deliberative reasoning among judges resembles multi-agent dialogue. By aligning these heterogeneous forms of reasoning, DiSCo–RAD demonstrates how the toolbox can be used in practice.

In our case study, the balancing layer captures how each judge weighs reasons of legality versus fairness, producing the dual-scale collaboration values of Table 1. The dialogue layer models interpersonal relations: formalist judges act as mutual supporters, while fairness-oriented judges form a separate alliance. The inference layer then builds a contrastive argumentation framework where each possible collaboration (e.g., "Judge α collaborates with Judge β") becomes an argument, and attacks represent incompatible choices. The semantics of this framework determine which collaborations are accepted and, through the coalition extraction function, which groups of judges form a rational and tolerant majority.

4 Formal Framework

In this section, we present two aligned routes from weighted pairwise relations to coalitions:

(1) Normative route (from social network $\rightarrow$ coalition): read *friend*/*enemy*/*frenemy* from the signs of pairwise collaboration values, then grow majority coalitions by a *friend-first + tolerance* closure satisfying IR and Tol(k).

(2) Argumentation route (contrastive BAF $\rightarrow$ extensions $\rightarrow$ coalition): encode directed *collaborations* as arguments; add a *local, intention-sensitive* attack and a *deductive* support relation that implements Tol(k); compute β-extensions [2]; extract coalitions from mutual-intent edges.

4.1 Balancing and social relations

We keep the balancing layer minimal.

Definition 4.1 [Weighted base and values] Let Ag be the set of agents and Ar the set of case reasons. A reason $(a, \alpha, \beta) \in R \subseteq Ar \times Ag \times Ag$ supports collaboration "α teams with β." A weighting $w : R \to \mathbb{R}^2_{\geq 0}$ attaches (w_1, w_2) for initiator/partner. Aggregation yields a *relative collaboration value* $\pi(\alpha, \beta) \in \mathbb{R}$ with sign indicating benefit (≥ 0) or loss (< 0) to α from teaming with β.

Definition 4.2 [Friend, enemy, frenemy] For $\alpha \neq \beta$:

$$\mathsf{friend}(\alpha, \beta) \iff \pi(\alpha, \beta) \geq 0 \ \wedge \ \pi(\beta, \alpha) \geq 0;$$

$$\mathsf{enemy}(\alpha, \beta) \iff \pi(\alpha, \beta) < 0 \ \wedge \ \pi(\beta, \alpha) < 0;$$

otherwise α, β are *frenemies*.

Tolerance is enforced via a simple margin inequality.

Definition 4.3 [Support with margin k] Given $k \geq 0$, β *supports* α *against* γ iff $\mathsf{friend}(\alpha, \beta) \ \wedge \ \pi(\alpha, \gamma) < 0 \ \wedge \ \pi(\alpha, \beta) + \pi(\alpha, \gamma) \ \geq \ k$.

4.2 Normative coalition formation: friend-first + tolerance

Let F be the undirected *friend graph* on Ag with edge $\{\alpha, \beta\}$ iff $\mathsf{friend}(\alpha, \beta)$. We grow coalitions from friend components and add frenemies only when bilateral tolerance is witnessed.

Definition 4.4 [Normative closure NORM_k] For each friend component S_0 of F, define the sequence

$$S_{i+1} = S_i \ \cup \ \{\gamma\}$$

where $\gamma \notin S_i$, and for every $\beta \in S_i$, $\mathsf{friend}(\gamma, \beta)$ or $\mathsf{frenemy}(\gamma, \beta)$ (there exist $\delta_1, \delta_2 \in S_i$ such that δ_1 supports γ against β or δ_2 supports β against γ).

Let $S^\star$ be the fixpoint of this iteration. A *normative coalition at margin* k is any $S^\star$ that also satisfies Maj (majority size).

By construction, any $S^\star$ produced by NORM_k satisfies IR (no enemy pair inside) and Tol(k) (frenemy pairs are bilaterally covered by allies); the Maj filter ensures feasibility.

4.3 Argumentation route: a contrastive bipolar AF

We now move to a bipolar argumentation framework (BAF) whose *arguments* are directed collaborations and whose relations encode *local attack* (exclusive choices from one agent's perspective) and *deductive support* (our tolerance rule).

Definition 4.5 [Contrastive BAF of collaborations] Fix $k \geq 0$. Let

$A := \{(\alpha, \beta) \in Ag \times Ag \mid \alpha \neq \beta\}$. Define the bipolar AF $\mathcal{B}_k = (A, \mathrm{Att}, \mathrm{Sup})$ by: $\mathrm{Att} := \{\, ((\alpha, \beta), (\alpha, \gamma)) \mid \pi(\alpha, \gamma) < 0 \,\}$ (*local, source-α attack*); and $\mathrm{Sup} := \{\, ((\beta, \alpha), (\alpha, \gamma)) \mid \mathsf{friend}(\alpha, \beta) \wedge \pi(\alpha, \gamma) < 0 \wedge \pi(\alpha, \beta) + \pi(\alpha, \gamma) \geq k \,\}$ (*deductive support implementing Tol(k)*).

Acceptability is defined via the dual *defeat/defence* for BAFs under the deductive reading of support: a set defeats an argument iff it defeats *all* its supporters; a set defends an argument iff it defeats *all* attackers of *some* supporter. These notions generalise Dung's AAFs and yield β-admissible/β-complete/β-preferred/β-stable/β-semi-stable semantics with standard properties (existence of β-complete/β-preferred, Fundamental Lemma, labelling correspondence). We rely on these definitions and results in what follows, see [2] for further details. [2]

Definition 4.6 [Defeat, defence, β-semantics] Let Sup^* be the reflexive and transitive closure of Sup. For $S \subseteq A$ and $a \in A$:

- S *defeats* a iff $\forall u \in \mathrm{Sup}^*(a)\ \exists b \in S :\ (b,u) \in \mathrm{Att}$;
- S *defends* a iff $\exists u \in \mathrm{Sup}^*(a)\ \forall b\big((b,u) \in \mathrm{Att} \Rightarrow S \text{ defeats } b\big)$.

Write $F_{\mathcal{B}}(S) := \{a \in A \mid S \text{ defends } a\}$. Then S is *conflict-free* iff it does not defeat any of its elements; S is β*-admissible* iff it is conflict-free and $S \subseteq F_{\mathcal{B}}(S)$. β-complete/β-preferred/β-stable/β-grounded/β-semi-stable extensions are defined as in AAFs but using $F_{\mathcal{B}}$ and defeat/defence above (e.g., β-preferred = maximal β-admissible).

4.4 Coalition extraction from β-extensions

A β-extension E collects the *possible collaborations*. We read coalitions from pairs that are mutually intended and then, if desired, prune back to IR.

Definition 4.7 [Mutual intent, maximal cliques, and extraction] Let $E \subseteq A$ be a β–extension. Define the *mutual–intent graph* $H_E = (Ag, E^{\leftrightarrow})$ with

$$E^{\leftrightarrow} := \big\{\{\alpha,\beta\} \subseteq Ag \;\big|\; (\alpha,\beta) \in E \ \wedge\ (\beta,\alpha) \in E\big\}.$$

A set $C \subseteq Ag$ is E*–endorsed* iff it induces a clique in H_E, i.e., $\{\alpha,\beta\} \in E^{\leftrightarrow}$ for all distinct $\alpha,\beta \in C$. Let $\mathsf{EXTClo}_k(E)$ be the family of *inclusion–maximal* E–endorsed sets that also satisfy **Maj**. (Optionally) apply an *IR–pruning* pass: remove endpoints that form enemy pairs inside C from one side; this does not break $\mathrm{Tol}(k)$ witnesses because supporters are friends by construction.

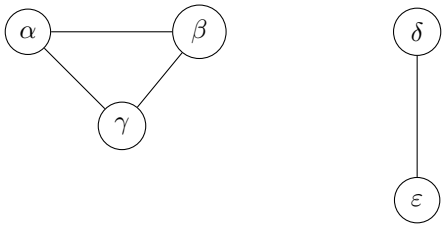

Fig. 4. Mutual–intent graph H_E: maximal cliques $\{\alpha,\beta,\gamma\}$ and $\{\delta,\varepsilon\}$. Only the triangle is a majority.

[2] We use the defeat/defence duality and β-semantics for Bipolar AFs in the sense of *deductive support*, which collapse to Dung's semantics when support is ignored [22,2].

5 Worked Illustration

We revisit the five-agent toy ($Ag = \{\alpha, \beta, \gamma, \delta, \varepsilon\}$) with collaboration values from Tables 1–2 and set $k = 1$.

Normative route. The friend graph has two components, $\{\alpha, \beta\}$ and $\{\delta, \varepsilon\}$. Starting at $S_0 = \{\alpha, \beta\}$, judge γ is a *frenemy* to both, but has bilateral coverage:

$$\pi(\alpha, \beta) + \pi(\alpha, \gamma) = 3 + (-2) = 1 \geq 1, \quad \pi(\beta, \alpha) + \pi(\beta, \gamma) = 2 + (-1) = 1 \geq 1.$$

Thus γ is added by NORM_1 (Def. 4.4), yielding $S^\star = \{\alpha, \beta, \gamma\}$, which satisfies IR, Tol(1), and Maj. The other component $\{\delta, \varepsilon\}$ cannot reach a majority.

Argumentation route. Build $\mathcal{B}_1 = (A, \mathrm{Att}, \mathrm{Sup})$ per Def. 4.5. For source α, every (α, β) attacks $(\alpha, \gamma), (\alpha, \delta), (\alpha, \varepsilon)$ because those targets have $\pi < 0$. For source β, $(\beta, \cdot)$ analogously attacks $(\beta, \gamma), (\beta, \delta), (\beta, \varepsilon)$. Deductive supports capture tolerance: (β, α) Sup (α, γ) since $\mathsf{friend}(\alpha, \beta)$ and $3 + (-2) \geq 1$; symmetrically, (α, β) Sup (β, γ) since $\mathsf{friend}(\alpha, \beta)$ and $2 + (-1) \geq 1$.

Under the β-semantics with dual defeat/defence, support neutralises local attacks precisely when supporters stand: to defeat (α, γ), one must defeat *all* its supporters in Sup^*; here, (β, α) is unattacked, so (α, γ) can be defended and accepted together with (α, β). The same holds for (β, γ) given (α, β). Hence a β-preferred (also β-complete) extension is:

$$E = \{(\alpha, \beta), (\beta, \alpha), (\alpha, \gamma), (\gamma, \alpha), (\beta, \gamma), (\gamma, \beta)\} \ \cup \ \{(\delta, \varepsilon), (\varepsilon, \delta)\}.$$

These β-semantics generalise Dung's AAFs and ensure existence and fixpoint properties under deductive support. [2]

Extraction and alignment. With the extension E from above, H_E has edges $\{\alpha, \beta\}, \{\alpha, \gamma\}, \{\beta, \gamma\}, \{\delta, \varepsilon\}$. The inclusion–maximal cliques are $\{\alpha, \beta, \gamma\}$ and $\{\delta, \varepsilon\}$; only the former satisfies **Maj**, hence $\mathsf{EXTClo}_1(E) = \{\{\alpha, \beta, \gamma\}\}$. Optional IR-pruning has no effect (no enemy pair inside). The coalition coincides with the normative one, so the alignment diagram commutes on this instance.

Remark (IR vs. Tol(k)). Because defeat requires eliminating *all* supporters, β-preferred sets can, in other instances, contain mutually negative pairs when sufficiently shielded by friends; EXTClo_k can then be followed by IR-pruning if strict IR is desired.

6 Related Work

Our approach follows the methodology promoted by Gabbay [23], he introduced the idea of building logics from reusable components—labels for time, agents, and resources; modalities for knowledge and obligation; and nonmonotonic rules—selected and combined to suit a given application. This view, known as *logic as a toolbox* [23,34], promotes methodological pluralism with engineering discipline: choose the right tools and make design choices explicit. Extending this vision, Gabbay and Rivlin promote argumentation as the core logic of interactive and explainable reasoning for the 21st century [24]. The

A–BDI metamodel [38] continues this trajectory by systematising formal argumentation into three complementary conceptualisations—balancing, dialogue, and inference—and showing how each can model different aspects of a legal case such as child custody. Complementarily, the RAD framework [1] aligns distinct reasoning routes and combining conceptualisations within one system. In this paper, we apply these methodological ideas to judicial discretion, integrating the three conceptualisations within a single toolbox model rather than using them in isolation.

Dung's abstract argumentation framework [22] is not only a generalisation of nonmonotonic inference but also unifies game-theoretic concepts. In his original paper, Dung observed that the stable extensions of an argumentation framework correspond to *von Neumann–Morgenstern stable sets* in cooperative game theory and to stable matchings in matching theory. This connection validates the use of argumentation semantics for modelling rational coalitions and equilibrium behaviour [3,17,14]. Building on this insight, our *contrastive argumentation framework* transfers the same stability intuition to the level of *directed collaborations* between agents, where local attacks and deductive supports define equilibrium-like acceptability dynamics. The resulting structure bridges abstract agent reasoning [39,10,9] and coalition formation [3,17,14], grounding DiSCo–RAD in both logical and social game semantics.

Arieli's work on reasoning under inconsistency [7,8] shares a similar idea with our conflict tolerance mechanism, which preserves classical acceptance while allowing controlled internal disagreement through a *support-with-margin* condition. Our model also draws on dual-scale balancing theories [35], representing benefit and cost for each collaboration. This approach complements broader work on argument strength [32,6], weighted argumentation [13,5], preference-based argumentation [29], and value-based reasoning [11]. Related extensions—gradual [12] and ranking-based semantics [4], probabilistic [28], and decision-theoretic models—explore how degrees of acceptability evolve under uncertainty and preferences.

From an application perspective, modelling *discretionary judicial reasoning* remains a major bottleneck for knowledge-representation and reasoning approaches. Only a few recent works explicitly address this challenge. Dik and Markovich's line of research formalises discretion as a *normative reasoning problem*, centred on the interplay between judicial freedom and its boundaries. Their deontic logic of discretion introduces *nuanced permissions* to represent degrees of judicial freedom in child-custody cases [19]. Building on this, their modal logic of the *duty of care* defines the *obligations* that constrain discretion—to determine all relevant factors, to weigh them properly, and to reason consistently [21]. In a subsequent work, they implement this normative characterisation in *Answer Set Programming* [20], modelling the judicial hierarchy, the obligation to be consistent, and the declaration of violations by higher-level courts [20]. We complement this research by modeling discretionary reasoning using *formal argumentation*. We focus on a specific and paradigmatic form of discretion—constitutional adjudication—and model how majority decisions

can emerge despite internal disagreement. In contrast to the deontic logic perspective, which describes the normative duties governing a single judge's reasoning, our framework captures the collective and relational dimension of discretion among judges. *DiSCo–RAD* thus operationalises discretionary reasoning as a form of *argumentative alignment*: a structured explanation showing how individual reasoning paths converge into a normatively acceptable collective outcome within the general methodology of formal and computational argumentation.

7 Summary and Outlook

This paper presented *DiSCo–RAD*, a Reasoning Alignment Diagram designed to explain how judicial majorities can emerge despite internal disagreement. Taking the Hungarian Constitutional Court as a case study, we modelled discretionary judicial reasoning through two complementary routes: a *normative route*, capturing coalition formation under rationality and tolerance postulates (IR, Tol(k), Maj), and an *argumentation route*, formalised as a contrastive bipolar framework with local attacks and deductive supports. When the two routes converge, the diagram provides an explanatory alignment between the normative and computational levels. Methodologically, the work illustrates the *logic-as-toolbox* idea and the A–BDI metamodel within a single framework. Formally, it introduces a contrastive acceptability model and a tolerance-aware coalition extractor that together account for how partial disagreement can coexist with collective rationality.

Outlook 1: methodological and technical mutual development of RAD. The design of the RAD can be viewed through two complementary lenses: the *methodological* and the *technical*. A useful precedent is the development of the ASPIC-family. On one hand, ASPIC provided a general modelling for structured argumentation; on the other hand, it was developed and refined with technical critiques and refinements—the systematic study of rationality postulates [16]—to ASPIC+ [30]. A similar separation clarifies our own setting. The methodological or meta-level contribution (the reasoning alignment between two routes) belongs in the methodological discussion, while the technical aspects—such as the precise attack design, the explicit treatment of defence and reinstatement, and the equivalence to direct coalition construction—belong in the technical part. This work remains in progress: closer interaction between the case study and the formal machinery will refine both. Some give and take on each side is not only natural but, in our view, necessary for eventual convergence between theory and application.

Outlook 2: combining formal argumentation with large language models. With the increasing capability of large language models (LLMs), a natural next step is to combine them with formal argumentation to simulate and analyse constitutional court cases. LLMs can generate, classify, and rephrase natural-language arguments, while formal argumentation provides the structure to evaluate and align them with normative standards. Such integration would extend the RAD methodology toward a hybrid reasoning

environment—linking symbolic and subsymbolic approaches—and bring the *logic-as-toolbox* vision closer to real-world applications. From an engineering perspective, this approach aligns with the theme of the workshop for *Logic for New-Generation AI*: selecting, combining, and standardising reasoning components that connect theoretical foundations with practical, explainable systems for applications.

Acknowledgments

We thank the anonymous reviewer for their comments. This work is supported by the Luxembourg National Research Fund (FNR) through the following projects: The Epistemology of AI Systems (EAI) (C22/SC/17111440), DJ4ME – A DJ for Machine Ethics: the Dialogue Jiminy (O24/18989918/DJ4ME), Logical Methods for Deontic Explanations (LoDEx) (INTER/DFG/23/17415164/LoDEx), Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN) (C24/19003061/SERAFIN), and the University of Luxembourg for the Marie Speyer Excellence Grant for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).

References

[1] *Special issue to celebrate dov gabbay's 80th birthday*, Journal of Applied Logics: The IfCoLog Journal of Logics and their Applications **12** (2025), pp. 1655–1685, to appear. Special Issue to Celebrate Dov Gabbay's 80th Birthday.

[2] Alcântara, J. and R. Cordeiro, *Bipolar argumentation frameworks with a dual relation between defeat and defence*, Journal of Logic and Computation **35** (2024).

[3] Amgoud, L., *An argumentation-based model for reasoning about coalition structures*, in: *International Workshop on Argumentation in Multi-Agent Systems*, Springer, 2005, pp. 217–228.

[4] Amgoud, L. and J. Ben-Naim, *Ranking-based semantics for argumentation frameworks*, in: *International Conference on Scalable Uncertainty Management*, Springer, 2013, pp. 134–147.

[5] Amgoud, L., J. Ben-Naim, D. Doder and S. Vesic, *Acceptability semantics for weighted argumentation frameworks*, in: *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, International Joint Conferences on Artificial Intelligence (IJCAI), 2017.

[6] Amgoud, L., D. Doder and S. Vesic, *Evaluation of argument strength in attack graphs: Foundations and semantics*, Artificial Intelligence **302** (2022), p. 103607.

[7] Arieli, O., *Conflict-tolerant semantics for argumentation frameworks*, in: *European Workshop on Logics in Artificial Intelligence*, Springer, 2012, pp. 28–40.

[8] Arieli, O., *Conflict-free and conflict-tolerant semantics for constrained argumentation frameworks*, Journal of Applied Logic **13** (2015), pp. 582–604.

[9] Arisaka, R., J. Dauphin, K. Satoh and L. van der Torre, *Multi-agent argumentation and dialogue*, IfCoLog Journal of Logics and Their Applications **9** (2022), pp. 921–954.

[10] Arisaka, R., K. Satoh and L. 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*, Springer, 2015, pp. 427–442.

[11] Atkinson, K. and T. J. Bench-Capon, *Value-based argumentation*, IfCoLog Journal of Logics and Their Applications **8** (2021), pp. 1543–1588.

[12] Baroni, P., A. Rago and F. Toni, *From fine-grained properties to broad principles for gradual argumentation: A principled spectrum*, International Journal of Approximate Reasoning **105** (2019), pp. 252–286.

[13] Bistarelli, S., F. Santini et al., *Weighted argumentation*, Handbook of Formal Argumentation, Volume 2 (2021).
[14] Boella, G., L. Van Der Torre and S. Villata, *Social viewpoints for arguing about coalitions*, in: *Pacific Rim International Conference on Multi-Agents*, Springer, 2008, pp. 66–77.
[15] Caminada, M., *Argumentation semantics as formal discussion*, Handbook of Formal Argumentation **1** (2018), pp. 487–518.
[16] Caminada, M. and L. Amgoud, *On the evaluation of argumentation formalisms*, Artificial Intelligence **171** (2007), pp. 286–310.
[17] Cayrol, C. and M.-C. Lagasquie-Schiex, *Coalitions of arguments: A tool for handling bipolar argumentation frameworks*, International Journal of Intelligent Systems **25** (2010), pp. 83–109.
[18] Dietrich, F. and C. List, *A reason-based theory of rational choice*, Nous **47** (2013), pp. 104–134.
[19] Dik, J. and R. Markovich, *Modeling judicial discretion with nuanced permissions*, in: *JURIX* (2024), pp. 48–59.
[20] Dik, J. and R. Markovich, *Judicial discretion as normative reasoning – deontic characterization of judicial decision making with answer set programming*, in: J. Maranhao, editor, *Proceedings of the 20th International Conference on AI and Law* (2025), pp. 258–267.
[21] Dik, J. and R. Markovich, *When judges go wrong: Modeling discretion and the duty of care*, in: *Deontic Logic and Normative Systems* (2025), pp. 399–400.
[22] Dung, P. M., *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*, Artificial intelligence **77** (1995), pp. 321–357.
[23] Gabbay, D. M., "Labelled deductive systems," Oxford university press, 1996.
[24] Gabbay, D. M. and L. 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).
[25] Gordon, T. F., *Towards requirements analysis for formal argumentation*, in: P. Baroni, D. Gabbay, M. Giacomin and L. van der Torre, editors, *Handbook of formal argumentation, Volume 1*, College Publictions, 2018 pp. 145–156.
[26] Henkin, L., *Infallibility under law: constitutional balancing*, Colum. L. Rev. **78** (1978), p. 1022.
[27] Hoek, W. v. d., L. B. Kuijer and Y. N. Wáng, *Who should be my friends? social balance from the perspective of game theory*, Journal of Logic, Language and Information **31** (2022), pp. 189–211.
[28] Hunter, A., S. Polberg, N. Potyka, T. Rienstra and M. Thimm, *Probabilistic argumentation: A survey*, Handbook of Formal Argumentation **2** (2021), pp. 397–441.
[29] Kaci, S., L. van der Torre, S. Vesic and S. Villata, *Preference in abstract argumentation*, in: D. Gabbay, M. Giacomin, G. R. Simari and M. Thimm, editors, *Handbook of Formal Argumentation, Volume 2*, College Publications, 2021 pp. 211–248.
[30] Prakken, H., *An abstract framework for argumentation with structured arguments*, Argument & Computation **1** (2010), pp. 93–124.
[31] Prakken, H., *Historical overview of formal argumentation*, in: *Handbook of formal argumentation*, College Publications, 2018 pp. 73–141.
[32] Prakken, H., *An abstract and structured account of dialectical argument strength*, Artificial Intelligence **335** (2024), p. 104193.
[33] Qiao, L., Y. Shen, L. Yu, B. Liao et al., *Arguing coalitions in abstract argumentation*, in: *Logics for New-Generation AI 2021*, CP College Publications, 2021 pp. 93–106.
[34] Tjitze, R., L. van der Torre and L. Yu, *Reasoning alignment for agentic ai: Argumentation, belief revision, and dialogue*, Journal of Applied Logics – IfCoLog Journal **12** (2025), pp. 1683–1712.
[35] Tucker, C., "The Weight of Reasons: A Framework for Ethics," Oxford University Press, 2025.
[36] Van der Hoek, W., L. Kuijer and Y. Wáng, *Logics of allies and enemies: A formal approach to the dynamics of social balance theory*, , **2020**, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 210–216.

[37] Yu, L., D. Chen, L. Qiao, Y. Shen and L. 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*, 2021, pp. 629–639.
URL `https://doi.org/10.24963/kr.2021/60`

[38] Yu, L. and L. 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)*, 2025, pp. 361–379.

[39] Yu, L., L. van der Torre and R. Markovich, *Thirteen challenges of formal and computational argumentation*, in: M. Thimm and G. R. Simari, editors, *Handbook of Formal Argumentation, Volume 3*, 2024 .