Where Do Weights Come From? Weighting and Weighing by LLMs and Formal Argumentation in Discretionary Judicial Reasoning

Davide Liga[1], **Réka Markovich**[1], **Liuwen Yu**[2]
[1]University of Luxembourg
[2]Luxembourg Institute of Science and Technology
{davide.liga,reka.markovich}@uni.lu, liuwen.yu@list.lu

Abstract

Discretionary judicial decisions depend on balancing competing considerations of various importance rather than mechanically applying norms. We examine whether large language models (LLMs) can perform reasonable *weight assignment* of the arguments used in judges' discretionary reasoning. Using a real child-custody case that proceeded through three judicial levels, we model each court's reasoning as a *bipolar argumentation* framework (BAF). The LLMs read the judges' reasoning in the sentences alongside with the corresponding BAF, and assign *numeric weights* to the arguments. We assess these assignments in two ways. First, a *robustness* analysis tests stability across repeated runs with identical prompts and decoding, asking whether the weights are consistent rather than noise. Second, we compute decisions from the resulting *weighted argumentation frameworks* (i.e., BAF with numeric weights from LLMs) and compare them with the courts' outcomes. In our case study, the assignments are stable across runs and the computed decisions reproduce the observed judicial pattern. We do not claim that assigning numbers amounts to a full account of legal reasoning; our aim is modest: to investigate LLMs' capacity in processing, grasping and reproducing discretionary judicial reasoning. Since this type is deemed by the legislator the ultimate human reasoning and decision making type of all those areas covered by the law, we see testing the capacities of—symbolic and subsymbolic—AI in this respect as investigating the frontiers of AI.

1. Introduction

Discretionary decision making of judges is a paradigmatic, still, particularly interesting form of legal reasoning under competing considerations. In the civil-law tradition, judicial reasoning is normally considered as a norm-based derivation, but certain domains—such as child custody—intentionally leave some kind of freedom to the judge, called *discretion*, because no single rule can determine the ideal outcome for all factual variations (Raz 1979; Shiner 1992; Alexy 2009; Hart 2012; Dworkin 2013). There is a(n intended) lack of ready-to-apply regulative norms. This freedom of the judge, however, is created by law and bounded by the *duty of care* (Dik and Markovich 2025a,b), which obliges judges to "examine carefully and impartially all relevant aspects of the individual case and to reason their decision accordingly". In our case study's area, Hungarian family law, the statute provides only the open-textured aim of securing "the best interests of the child" in child custody cases. Based on case law confirmed by Hungarian Supreme Court's Directive 17, the judge has to fulfill the duty of care by weighing heterogeneous factual grounds—emotional bond, stability, cooperation, schooling, and others—and assign them relative importance (Dik and Markovich 2024). Discretionary judicial decision making is therefore an exercise in a balancing model: it is neither mechanical rule-application nor arbitrary choice, but a structured comparison of diverse reasons within normative limits. However, judges never use numerical weights to express an argument's importance, so when one tries to model their decisions, one has to answer the question: Where do weights come from? Since the written judicial sentence actually constitutes the official reasongiving of a discretionary decision, relying on its text is not only our sole option, but also the best one when assigning weights for formal reconstruction.

Against this background, we test whether LLMs can *weight reasons* in a real discretionary case by *reverse-engineering* the judges' reasoning (Tucker 2025): "when a decision is certain, we can reverse engineer the (embedded) weights of reasons: we know which acts are permissible and which aren't, we identify the weights of reasons that would explain that pattern of reasoning, and then we infer that those weights obtain." Concretely, we encode each judicial level's reasoning as a BAF (Cayrol and Lagasquie-Schiex 2009), which naturally represents pro and con arguments and their interactions, and we prompt LLMs to read the judges' descriptions (e.g., "decisive factor," "over-weighted," "underweighted") and assign *numeric base strengths* to the argument nodes. We then evaluate the assignments in two complementary ways. First, a robustness check: across repeated runs with identical prompts/decoding, do node weights and their induced rankings/acceptability labels remain stable (i.e., more than random noise)? Second, an outcome-alignment check: using the LLM-assigned weights to instantiate *weighted bipolar argumentation frameworks* (WBAF) (Amgoud et al. 2017), do the decisions computed from the WBAF *give the actual verdicts* at each judicial level? This hybrid setup probes whether subsymbolic models exhibit context-sensitive factor weighting when embedded in a clear symbolic structure, and whether such weight-

ing is both *internally stable* and *externally aligned* with the courts' outcomes. The results show the weight assignment produced by the LLM is consistent. While this confirms that the numerical weights generated by the LLM are not just random noise, we are not claiming that LLMs' reasoning is equal (nor can replace) human reasoning. We only claim that: (a) LLMs' weight assignments show both internal and external consistency, and (b) this consistency appears aligned with human reasoning.

The layout of this paper is as follows. Section 2 introduces the child custody case that serves as our running example. Section 3 presents the formal preliminaries of BAF and WBAF used to model the judges' reasoning. Section 4 describes the experimental setup in which LLMs assign weights to the arguments and evaluates the consistency of these weights. Section 5 applies the formal semantics to the WBAF to compute final decisions and compare them with the judicial outcomes. Section 6 reviews related work, and Section 7 concludes with future directions.

2. Case Background

The case considered in this paper concerns a child custody dispute in Hungary [1]. The issue was whether custody should remain with the mother, with whom they had continuously lived since separation, or be transferred to the father. Both parents satisfied objective living conditions, but the courts diverged in how they *weighed* subjective factors: continuity of care, stability, parental suitability, the children's wishes, and the alleged "alienating" conduct of the mother. The Supreme Court annulled the appellate decision and restored the first-instance outcome, granting custody to the mother. For formal modeling, we assign symbolic labels to arguments. Let A denote the decision "custody to the mother" and B the decision "custody to the father." Arguments supporting (resp. attacking) A are labeled Ap1, Ap2, ... (resp. Ac1, Ac2, ...), and analogously Bp1, Bp2, ... and Bc1, Bc2, ... for B. Subscripts (e.g., Ap4p1) denote higher-level arguments targeting other arguments rather than the decision node. This notation systematically links legal reasoning to its representation in Weighted Bipolar Argumentation Frameworks (WBAFs).

2.1. Judge 1 (first instance)

Decision: Custody to the mother (A).
Pros for the mother (supporting A):

- **Ap1:** Mother can provide objective conditions (housing, care).
- **Ap2:** Children have a close emotional bond with Mother; loving environment.
- **Ap3:** Mother's educational suitability is better (expert opinion).
- **Ap4:** Children's wishes; want to live with Mother.
- **Ap4p1:** Older child's consent rule (legal significance of §74 Csit).
- **Ap4p2:** Children are of sufficient age (15 and 13) and maturity to have their wishes respected.
- **Ap4p3:** Siblings' bond / no separation should be preserved.
- **Ap5:** Children have settled and integrated in Germany.

Cons for the mother (attacking A):

- **Ac1:** Mother did not promote contact with Father; was overprotective.
- **Ac2:** Non-consensual move to Germany.
- **Ac2p1:** Non-consensual move is illegal under Hague Art. 3.
- **Ac2p1c1:** Hague Convention Art. 13(2); mature children's objections justify refusal of return.

Pros for the father (supporting B):

- **Bp1:** Father can provide objective conditions (housing, care).
- **Bp2:** Younger child strongly attached to Father.

Cons for the father (attacking B):

- **Bc1:** Father involved the children in parental conflict.
- **Bc2:** Poor relationship with the older child.
- **Bc2p1:** Violent incident (spoon hitting) against the older child.

2.2. Judge 2 (second instance)

Decision: Custody to the father (B).
Pros for the mother (supporting A):

- **Ap1:** Mother can provide objective conditions (housing, care).
- **Ap2:** Mother's educational suitability rated average (not stronger than Father).
- **Ap3:** Children's wishes; want to live with Mother.

Cons for the mother (attacking A):

- **Ac1:** Non-consensual move to Germany.
- **Ac2p1:** Non-consensual move is illegal under Hague Art. 3.
- **Ac1p2:** Non-consensual move shows Mother alienated the children from Father.
- **Ac1p3:** Mother exposed children to repeated environmental changes / instability (moves to guesthouse, Austria, then Germany).
- **Ap3c1:** Mother's alienating conduct manipulates children's wishes.

Pros for the father (supporting B):

- **Bp1:** Father can provide objective conditions (housing, care).
- **Bp2:** Younger child strongly attached to Father.

Cons for the father (attacking B):

- **Bc1:** Poor relationship with the older child.

Supreme Court

Decision: The appellate decision was annulled; custody was restored to the mother.
Reasoning:

- The appellate court over-weighted alienating conduct and under-weighted continuity and the children's wishes.
- The wishes of older minors could not be dismissed as manipulation.
- Both parents had comparable suitability, but continuity and the children's will tipped the balance.

[1] Hungarian Supreme Court (Kúria), *Partial Judgment*, Case No. Pfv.II.20.900/2021/13.

Judges may weight pros and cons differently. We make this explicit by modeling the reasoning as a Bipolar Argumentation Framework (BAF), where pros and cons are argument nodes and arrows represent support and attack. This enables quantitative analysis in Section 3.

3. Model the Case with Weighted Bipolar Argumentation

Discretionary judicial reasoning can be seen as a process of *balancing reasons*, where factual and normative considerations provide *support* for or *attack* against competing legal conclusions. To make this structure explicit, we use *weighted bipolar argumentation* (Amgoud and Ben-Naim 2018). Abstract argumentation (Dung 1995) treats each relevant consideration as an *argument* (a node) and the relations among arguments as *interactions* (edges). A *bipolar* framework distinguishes two interaction types—*support* and *attack*—which correspond directly to the *pros* and *cons* that judges weigh in deliberation. Weighted bipolar argumentation adds a numerical base weight to each argument, capturing its initial plausibility, credibility, or importance. This quantitative layer is particularly suitable for the custody dispute examined here, where judges do not disagree about the set of considerations but about the *relative importance* they assign to them.

We first represent each judge's reasoning as a *bipolar argumentation framework* (BAF) (Cayrol and Lagasquie-Schiex 2009), i.e., *without* weights.

Definition 1 (Bipolar Argumentation Framework)**.** A *Bipolar Argumentation Framework (BAF)* is a triple $\mathcal{A} = \langle A, R, S\rangle$, where A is a finite set of *arguments*, $R \subseteq A \times A$ is the set of *attack* relations (with $(b, a) \in R$ meaning b attacks a), and $S \subseteq A \times A$ is the set of *support* relations (with $(b, a) \in S$ meaning b supports a).

Figures 1 and 2 visualise Judge 1 and Judge 2, respectively. In both figures, *solid arrows* denote **attacks** and *dashed arrows* denote **supports**. These diagrams fix the *structure* (arguments and support/attack relations) independently of any numerical assessment.

We now extend BAF with base weights to the same argument sets to obtain *weighted* BAFs that have been defined in (Amgoud et al. 2017).

Definition 2 (Weighted Bipolar Argumentation Framework)**.** A *Weighted Bipolar Argumentation Framework (WBAF)* is a quadruple $\mathcal{A} = \langle A, w, R, S\rangle$, where $\langle A, R, S\rangle$ is a BAF and $w : A \to [0, 1]$ assigns each argument $a \in A$ its *intrinsic strength*.

Notation. For $a \in A$, let $\mathrm{Supp}(a) = \{x \in A \mid xSa\}$ and $\mathrm{Att}(a) = \{x \in A \mid xRa\}$. We write WAG for the set of all weighted bipolar graphs $\langle A, w, R, S\rangle$.

We fix the topology (A, R, S) from Section 2, as visualized in Figures 1–2, and obtain base weights $w : A \to [0, 1]$ from LLM prompts. Decision nodes A and B are anchored at 0.5; only non-decision nodes vary across runs.

We evaluate acyclic, non-maximal weighted graphs with a *restricted* semantics.

Definition 3 (Restricted semantics (Amgoud et al. 2017))**.** A *restricted semantics* is a function S that maps any acyclic, non-maximal $\mathcal{A} = \langle A, w, R, S\rangle \in WAG$ to a function $\mathrm{Deg}^S_{\mathcal{A}} : A \to [0, 1]$.

We work with acyclic graphs (the relation $R \cup S$ induces no directed cycle) and *non-maximal* graphs (no node has weight 1). Our case maps are DAGs (pros/cons flow toward the decision nodes).

Definition 4 (Well-founded relation)**.** Let $\mathcal{A} = \langle A, w, R, S\rangle$ be an acyclic WBAF and $a \in A$. A *path to* a is a non-empty finite sequence $\langle a_1, \ldots, a_n\rangle$ with $a_n = a$ and $\langle a_i, a_{i+1}\rangle \in R \cup S$ for each $i < n$. Define a well-founded relation $\prec$ on A by $x \prec y$ iff the maximum length of a path to x is strictly smaller than that to y.

We adopt the Euler-based restricted semantics (Ebs) (Amgoud et al. 2017) to compute an acceptability degree $f(a) \in [0, 1]$ for each argument from its initial weight and the aggregate effect of its attackers and supporters.

Definition 5 (Euler-based restricted semantics (Amgoud et al. 2017))**.** Let $\mathcal{A} = \langle A, w, R, S\rangle$ be acyclic and non-maximal. The *Ebs* semantics is the function $f : A \to [0, 1]$ defined recursively along $\prec$ by

$$f(a) \;=\; 1 - \frac{1 - w(a)^2}{1 + w(a)\, e^{E}}, E \;=\; \sum_{x \in \mathrm{Supp}(a)} f(x) - \sum_{x \in \mathrm{Att}(a)} f(x).$$

The overall degree of a is $\mathrm{Deg}^{\mathsf{Ebs}}_{\mathcal{A}}(a) = f(a)$.

With the structure fixed (BAF), the base weights (WBAF), and a semantics in place (Ebs), the next section tests whether the resulting WBAFs reproduce the judges' outcomes and how stable the LLM-derived weights are across runs.

4. Experiment and Evaluation

We construct two WBAFs from the case: **J1** (first judge) and **J2** (second judge), each with decision nodes **A** (custody to mother) and **B** (custody to father). The *structure* (nodes and typed edges) is fixed across runs; only node weights vary. Excluding the decision nodes A and B, J1 has 17 nodes and 17 edges, while J2 has 11 nodes and 12 edges. Counting A/B, J1 has 19 nodes and 19 edges; J2 has 13 nodes and 14 edges. The two graphs share some nodes and edges but are different graphs which reflect the specific judge's legal reasoning.

Our experiment is divided into two tasks, T1 and T2. T1 tests the ability of LLMs to give weights to the two Weighted Bipolar Argumentation graphs (J1/J2) consistently. To evaluate this first task, we checked whether the LLM outputs converge when performing K runs. T2 tests the ability of LLMs to give an overall judgement (simulating the Supreme Court's process of evaluating the views from the first and second instance). In other words, the idea is to generate J3 as a Weighted Bipolar Argumentation graph. Also in this case, as evaluation, we checked whether the LLM outputs converge when performing K runs.

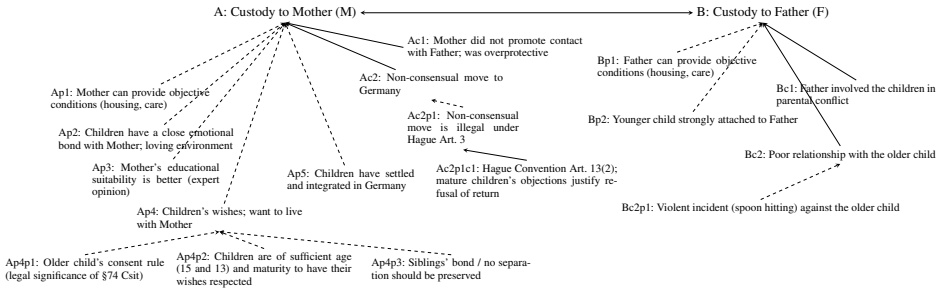


Figure 1: Reasoning of judge 1

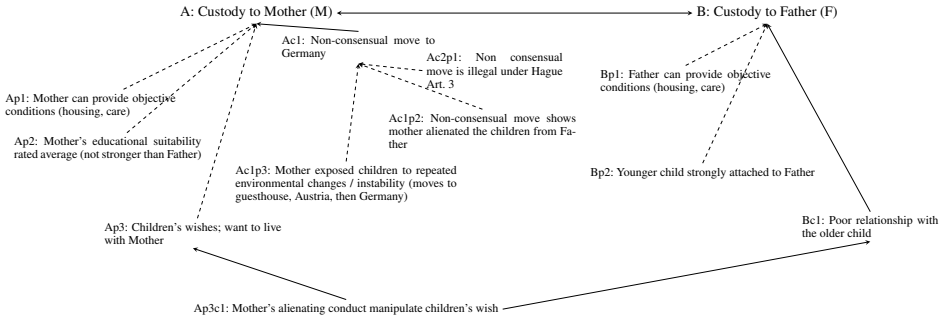


Figure 2: Reasoning of judge 2

4.1. Task 1 (weighting J1 and J2)

In T1, we prompted the LLM(s) to generate $K{=}10$ independent runs for each graph (J1/J2) by re-prompting the model with identical content and decoding settings and asking the LLM to assign weights to each node in J1 and J2.

Prompting protocol. We provide the **full graph structure**: the list of nodes and their IDs, and the typed edges (support/attack). We ask the model to return only the **node weights** (numbers in $[0, 1]$) in JSON, keeping the edges unchanged. We *instruct* the model that the two decision nodes **A** (Mother) and **B** (Father) should be 0.5. The following prompt template has been used across all runs (for brevity, we reported some placeholders in bold).

A child custody case was examined by two different judges who assessed who should have custody of the child between the mother and the father. There are two possible main claims:

- A: Custody to Mother
- B: Custody to Father

The judges analyzed and presented the pros and cons of these two opposing claims. Below are the pros (p) and cons (c) provided by the two judges (and their relative justifications). Each pro/con has a unique name (e.g., "Ap1" means "pro 1 to A"; "Ap4p1" means "pro 1 to pro 4 of A").

YOUR TASK
– According to JUDGE 1: –
{Pros/Cons to A according to Judge 1}
{Pros/Cons to B according to Judge 1}
{Attack/Support relations according to Judge 1}

– According to JUDGE 2: –
{Pros/Cons to A according to Judge 2}
{Pros/Cons to B according to Judge 2}
{Attack/Support relations according to Judge 2}
Your task is to assign weights to each one of the above elements, therefore composing a *Weighted Bipolar Argumentation Framework* for each judge. A Bipolar Argumentation Framework (BAF) is a directed graph whose nodes are arguments, and whose edges can be of two types: attack edges and support edges. A Weighted Bipolar Argumentation graph extends this structure by associating a numerical weight, ranging from 0 to 1, with each node. In this task, you should assign weights to the nodes. The final decision nodes A and B must always have an equal and neutral weight of 0.5.

Format your output as follows:

```
{"nodes": [ { "id": "A",    "weight": ... },
            { "id": "B",    "weight": ... },
          { "id": "Ap1",  "weight": ... }, ...]}
```

We executed $K = 10$ runs with identical prompt and decoding settings.

4.2. Task 2 (generating J3)

Task 2 closely mirrors Task 1 but asks the LLM to *generate* a new Weighted Bipolar Argumentation Framework (**J3**) that simulates the Supreme Court's reasoning. The model receives the two lower-court structures and is instructed to *synthesize* a final graph representing the Supreme Court's evaluation. The topology is not hand-merged: the LLM proposes the node set and typed edges (support/attack) following the same labeling scheme (A/B, Ap/Ac, Bp/Bc). Decision nodes

remain anchored at 0.5, as in Task 1. The prompt is identical to Task 1 except for the "Your Task" portion, which was modified as follows (change in **bold**):

> ...
> YOUR TASK
> **Your task is to generate a single Weighted Bipolar Argumentation Framework that provides a final legal decision, simulating the role of the Supreme Court over the two lower-instance judges.** A Bipolar Argumentation Framework (BAF) is a ...

4.3. Evaluation

To measure the quality of the output for both Task 1 and Task 2, we first designed a simple formula to obtain an *outcome score* for each node and then used such scores in our *evaluation metrics*.

Outcome Score To achieve this first outcome score:

1. For a node v, sum the weights of all **incoming supporters** and subtract the weights of all **incoming attackers**.
2. Scale that net influence by a value α (which indicates how much the graph structure matters).
3. Add the influence score from step 2 to v's own weight.
4. Clip the result element-wise to $[0, 1]$ (values below 0 become 0; above 1 become 1). This clipping is independent of the decision threshold τ.

Formally,

$$s \;=\; \mathrm{clip}_{[0,1]}\big(w + \alpha\, E^{\top} w\big),$$

where $E \in \{-1, 0, +1\}^{n\times n}$ is the signed adjacency with $E_{uv} = +1$ if $u \to v$ is support, $E_{uv} = -1$ if attack, and 0 otherwise. Using $E^{\top}w$ aggregates *incoming* influence for each node. We call s the outcomes because they are the weights after accounting for direct supporters and attackers; we call the operation a one-step influence update because each node aggregates incoming signed influence once (no iterative propagation). We report results at $\alpha \in \{0.3, 0.5, 0.7\}$.

Decision rule. A node is *accepted* if its outcome $s \geq \tau$ (default $\tau = 0.5$). We also show how results change when τ varies. Note that τ affects only acceptance decisions; outcomes are always clipped to $[0, 1]$ independently of τ. Clipping enforces probabilistic bounds on s; τ only binarizes s for decision-level agreement metrics.

Mini example. Suppose node X has base weight $w_X = 0.6$, with one supporter Y ($w_Y = 0.8$) and one attacker Z ($w_Z = 0.5$). With $\alpha = 0.5$:

$$s_X = \mathrm{clip}_{[0,1]}\big(0.6 \;+\; 0.5 \cdot (0.8 - 0.5)\big) \quad = \mathrm{clip}_{[0,1]}(0.75) \\ = 0.75.$$

Since $0.75 \in [0, 1]$, clipping does not apply (clipping would truncate to the nearest bound only if the value were < 0 or > 1). After clipping, the decision rule is applied: with $\tau = 0.5$, $s_X \geq \tau$ and X is accepted.

We treat α (propagation influence) and τ (acceptance cutoff) as *analysis hyperparameters* for sensitivity studies; they are not learned and do not affect the raw weight outputs.

Metrics We first compute per-run *outcomes* $s = \mathrm{clip}_{[0,1]}(w + \alpha E^{\top} w)$ (Section 4.3). These outcomes reflect each node's base weight *after* one step of signed support/attack influence. We then assess stability and reliability along three complementary axes.

Legend: ↑ means "higher is better"; ↓ means "lower is better".

(A) Continuous agreement (shape & magnitude). These metrics treat runs as real-valued vectors and ask: *do different runs place **weight** on the same nodes, and by how much do they differ?*

- **Pairwise cosine** (↑): directional alignment of outcome vectors across all run pairs; near 1 means runs agree on the *relative configuration* of nodes, regardless of scale.
- **Pairwise RMSE** (↓): average magnitude of differences across run pairs; lower means tighter clustering in absolute terms. (We compute L1 as well; results are consistent.)
- **To-mean cosine / RMSE** (↑/↓): agreement of each run with the *ensemble mean* (element-wise average), then averaged over runs. This shows how concentrated runs are around a single "consensus" point.

Unless noted, we report these on outcomes s (they can also be computed on weights w; we include w in node-level diagnostics below).

(B) Reliability across runs (rank- & variance-based). We treat nodes as subjects and runs as raters and ask: *are nodes ordered consistently across runs, and how much of the total variance is due to node differences rather than intra-run noise?*

- **Kendall's** W (↑): rank concordance of nodes across runs; $W = 1$ means identical rankings (ties handled by averaged ranks).
- **ICC(2,k)** (↑): two-way random-effects, absolute-agreement ICC for the average of k runs. Intuition: do runs produce the same numeric values (no systematic offsets/scale differences), so that the mean across runs is reliable?
- **ICC(3,k)** (↑): two-way mixed-effects, consistency ICC for the average of k runs. Intuition: do runs preserve relative differences between nodes even if each run has a fixed additive bias (e.g., one run is uniformly a bit higher/lower)?

Values near 1 indicate that most variability is between nodes rather than between runs, i.e., the run process is stable.

(C) Decision-level agreement (accept/reject). We binarize outcomes using the threshold τ and ask: *do runs make the same discrete choice?*

- **Fleiss'** κ (↑): multi-run agreement on labels $\mathbb{1}[s \geq \tau]$ beyond chance. We also report the mean acceptance rate per node for context.

Node-level diagnostics (where variability lives). For each task (J1, J2, J3) we visualize, across *all models*, the **top outcome ranges** (max–min of s across the 10 runs), the **top weight ranges** (max–min of w), and the **acceptance rates**

Table 1: Inter-run stability at α=0.5, τ=0.5 on outcomes s. Brackets show sensitivity ranges: for Cos and RMSE, Δ_α over $\{0.3, 0.5, 0.7\}$; for κ, Δ_τ over $\{0.45, 0.50, 0.60\}$. ↑ higher is better; ↓ lower is better. Green numbers refer to best values; orange numbers refer to worst values.

Task	Cos(s)↑ [Δ_α]	RMSE(s)↓ [Δ_α]	W ↑	ICC(2,k)/(3,k)↑	κ ↑ [Δ_τ]
		GPT-5			
J1(T1)	0.9965 [0.0005]	0.0638 [0.0063]	0.9211	0.9948 / 0.9951	0.8180 [0.1820]
J2(T1)	0.9957 [0.0010]	0.0668 [0.0067]	0.9376	0.9879 / 0.9881	0.9090 [0.0952]
J3(T2)	0.9947 [0.0024]	0.0760 [0.0234]	0.9197	0.9935 / 0.9938	0.9395 [0.0228]
		DeepSeek			
J1(T1)	0.9898 [0.0001]	0.1013 [0.0023]	0.8519	0.9831 / 0.9843	0.7033 [0.1146]
J2(T1)	0.9956 [0.0007]	0.0599 [0.0027]	0.9166	0.9949 / 0.9947	1.0000 [0.0000]
J3(T2)	0.9969 [0.0004]	0.0594 [0.0027]	0.9429	0.9963 / 0.9964	0.9248 [0.0936]
		Claude			
J1(T1)	0.9991 [0.0001]	0.0345 [0.0010]	0.9592	0.9984 / 0.9986	1.0000 [0.0000]
J2(T1)	0.9978 [0.0008]	0.0480 [0.0075]	0.9447	0.9969 / 0.9971	1.0000 [0.1816]
J3(T2)	0.9964 [0.0010]	0.0596 [0.0043]	0.9527	0.9953 / 0.9956	0.7645 [0.2021]

Notes: DeepSeek model: DeepSeek-V3.2-Exp; Claude model: Claude Sonnet 4.5.
Ranges: $\Delta_\alpha = \max_{\alpha\in\{0.3,0.5,0.7\}} - \min_\alpha$;
$\Delta_\tau = \max_{\tau\in\{0.45,0.50,0.60\}} - \min_\tau$.

at τ=0.5. These grouped bar plots (Figs. 3–5) show, for each node, side-by-side bars for the different models—making it easy to see which nodes drive disagreement and which models are tighter or looser on those nodes.

Sensitivity analyses (robustness to α, τ). We treat α (propagation influence) and τ (acceptance cutoff) as analysis hyperparameters. We sweep $\alpha \in \{0.3, 0.5, 0.7\}$ and $\tau \in \{0.45, 0.50, 0.60\}$; we report max–min ranges Δ_α (Cos/RMSE) and Δ_τ (κ) directly in Table 1.

Qualitative labels (readability only). For scalar metrics, we attach *excellent/good/fair/poor* tags using fixed, task-agnostic cutoffs to aid quick reading; these tags do not affect computations.

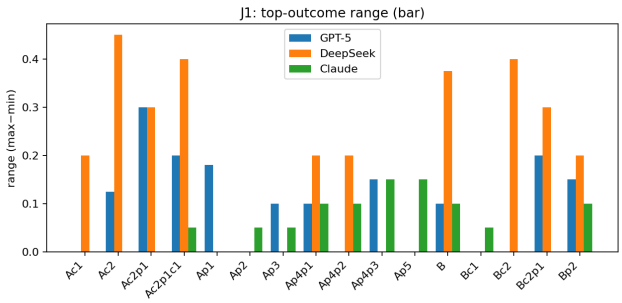
Reporting note. Table 1 consolidates headline stability for every *model* × *task*, including inline sensitivity ranges: Δ_α for Cos/RMSE (over $\alpha \in \{0.3, 0.5, 0.7\}$) and Δ_τ for κ (over $\tau \in \{0.45, 0.50, 0.60\}$). Node-level diagnostics are shown for each task—across all models—in Figs. 3, 4, and 5.

Evaluation Protocol For each graph (J1/J2) in T1 and for J3 in T2, we:

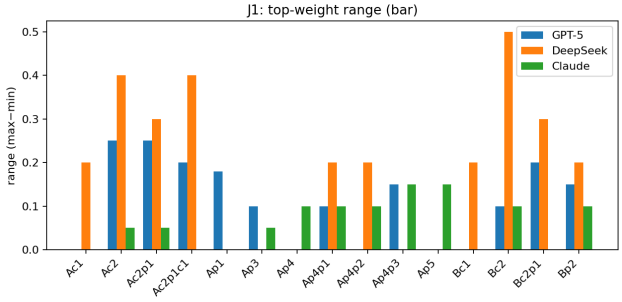
1. **Compute outcomes.** From each run's weights w, compute $s = \text{clip}_{[0,1]}(w + \alpha E^\top w)$ at the default α=0.5.
2. **Headline stability & reliability.** Evaluate continuous agreement (pairwise and to-mean cosine/RMSE), reliability (Kendall's W, ICC(2,k), ICC(3,k)), and decision-level agreement (Fleiss' κ at τ=0.5). Report per model/task in Table 1.
3. **Sensitivity (inline).** Repeat headline metrics for $\alpha \in \{0.3, 0.5, 0.7\}$ and $\tau \in \{0.45, 0.50, 0.60\}$; record the resulting ranges Δ_α and Δ_τ in Table 1.
4. **Diagnostics.** Plot grouped bar charts of per-node variability (top outcome range, top weight range) and acceptance rates across models; see Figs. 3–5.

4.4. Evaluation Takeaways

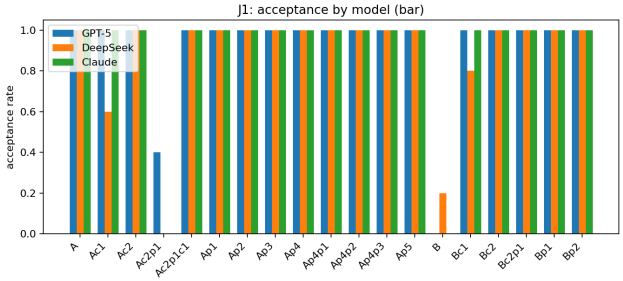
Strong stability across runs. Outcome vectors are highly aligned (cosine ≈ 1) with small dispersion (low RMSE), and



(a) Outcome range (max–min)



(b) Weight range (max–min)



(c) Acceptance rate

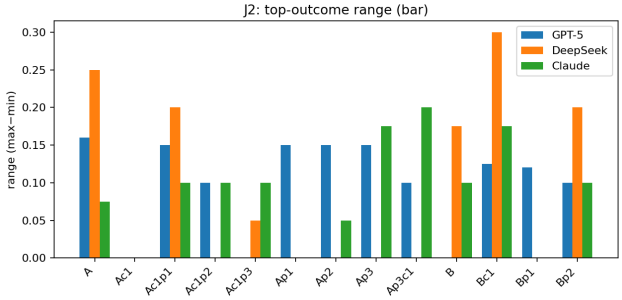
Figure 3: Task 1 (J1): node-level diagnostics across runs by model (GPT-5, DeepSeek, Claude). Bars within each node compare models; acceptance at τ=0.5.

each run lies close to the ensemble mean, indicating tight clustering.

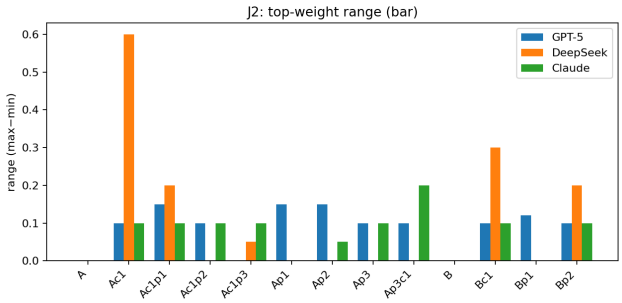
High reliability and consistent decisions. Both rank-based (Kendall's W) and variance-based (ICC(2,k), ICC(3,k)) measures are near 1, showing that node ordering and absolute levels are consistent across runs. Agreement on accept/reject labels (Fleiss' κ) is *excellent* at τ=0.5, and remains high across reasonable τ choices.

Robust to α, τ. Headline metrics change only mildly across the tested propagation influences and thresholds, suggesting conclusions are not sensitive to these settings.

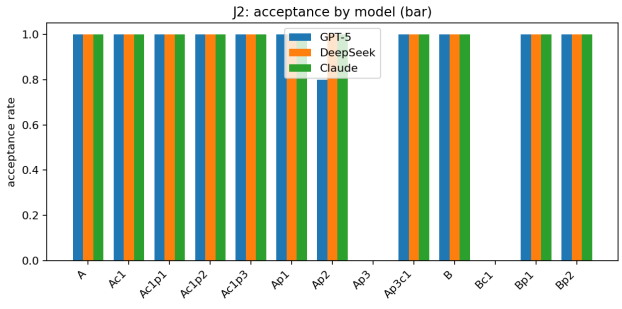
Actionable diagnostics. Variability concentrates on a small set of second-order/meta nodes, guiding targeted prompt or calibration adjustments; decision nodes remain stable across runs given the instruction to set A and B to 0.5 and the evaluation uses outcomes clipped to $[0, 1]$ as defined in Section 4.3.



(a) Outcome range (max–min)

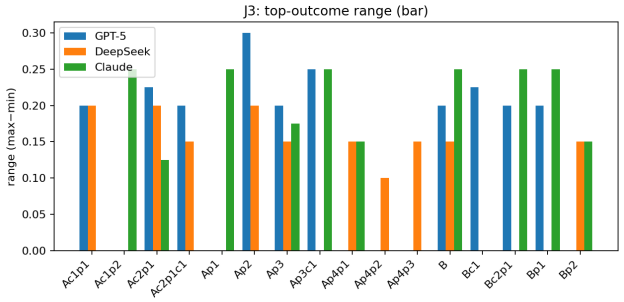


(b) Weight range (max–min)

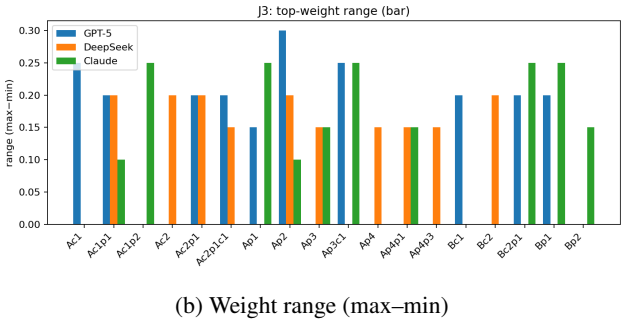


(c) Acceptance rate

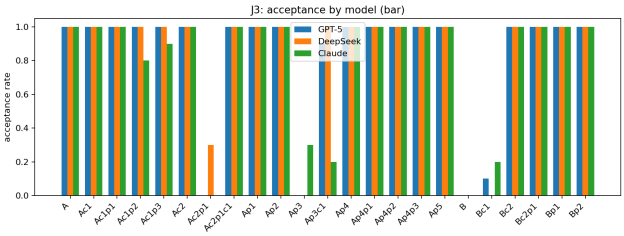
Figure 4: Task 1 (J2): node-level diagnostics across runs by model (GPT-5, DeepSeek, Claude).



(a) Outcome range (max–min)



(b) Weight range (max–min)



(c) Acceptance rate

Figure 5: Task 2 (J3): node-level diagnostics across runs by model (GPT-5, DeepSeek, Claude).

4.5. Discussion on Weight Assignment Results

Overall stability. Across all models and tasks, inter-run variability is low: cosine similarities are near 1 and RMSEs are small, while rank- and variance-based reliability (Kendall's W, ICC(2,k), ICC(3,k)) are consistently high (Table 1). Sensitivity to the propagation parameter is uniformly minor: the Δ_α ranges in the Cos/RMSE columns are tiny, indicating conclusions are not artifacts of a particular α.

Task 1 (J1/J2): Claude is the most stable model. It delivers the best continuous and reliability scores on both J1 and J2 (*bold green* entries in Table 1), with perfect decision agreement at τ=0.5 (κ=1.0). GPT-5 is close behind on J1 but lags on J2 (with the 'worst' RMSE and ICCs for GPT-5 on J2). DeepSeek shows a split: it is weakest on J1 across all metrics (**Cos**, **RMSE**, W, **ICC**, κ), yet achieves **perfect** decision robustness on J2 (κ=1.0 with Δ_τ=0), meaning its accept/reject labels are invariant to the tested thresholds even if its continuous dispersion is not the best one.

Task 2 (J3): The "Supreme Court" synthesis is more demanding. DeepSeek attains the best continuous stability (best Cos/RMSE and ICCs), while Claude attains the best rank concordance (W). GPT-5 shows the weakest continuous stability on J3 (**Cos**, **RMSE**, W, **ICC**), but it yields the **best** decision-level agreement at τ=0.5 (highest κ) and the smallest Δ_τ among the three, indicating comparatively robust binary decisions once the threshold is set. In contrast, Claude's J3 decisions are the most threshold-sensitive (lowest κ and largest Δ_τ), despite its excellent continuous/rank behavior.

Practical interpretation. If the downstream goal is *stable continuous weighting or rankings* (e.g., for explanatory analyses or aggregation), Claude is a strong default for T1, and DeepSeek is preferable for J3. If the pipeline hinges on *stable binary accept/reject decisions*, one should favor configurations with high κ and small Δ_τ—notably Claude on J1, DeepSeek on J2, and GPT-5 on J3.

5. Semantic Evaluation and Final Decisions

Using the averaged node weights from ten runs of each LLM (Claude 4.5, GPT-5, and DeepSeek), we computed the overall acceptability of each argument with the *Euler-based restricted semantics (Ebs)* defined in Section 3. This semantics propagates the influence of supporting and attacking arguments through the weighted graph to yield final acceptability scores $f(a)$ for all nodes. The two decision arguments, A (*custody to mother*) and B (*custody to father*), were anchored at 0.5, and their resulting scores determine the outcome: if $f(A) > f(B)$, the model predicts custody to the mother; otherwise, to the father.

Table 2 reports the computed acceptability of the decision nodes for all models and graphs. The last column indicates which side the weighted semantics favors.

Table 2: Final decisions computed with Ebs semantics (higher value indicates preferred decision).

Graph	Model	$f(A)$ (Mother)	$f(B)$ (Father)	Decision
J_1 (First Instance)	Claude 4.5	0.855	0.482	Mother
	GPT-5	0.857	0.466	Mother
	DeepSeek	0.852	0.517	Mother
J_2 (Appeal)	Claude 4.5	0.549	0.658	Father
	GPT-5	0.523	0.652	Father
	DeepSeek	0.518	0.666	Father
J_3 (Supreme Court)	Claude 4.5	0.813	0.515	Mother
	GPT-5	0.768	0.513	Mother
	DeepSeek	0.743	0.512	Mother

The application of the Euler-based semantics to LLM-assigned weights successfully reproduces the pattern of judicial outcomes across all levels. In the first-instance graph (J_1), all models favor custody to the mother; in the appellate graph (J_2), the preference shifts to the father; and in the Supreme Court synthesis (J_3), the evaluation again favors the mother—mirroring the respective court decisions. These results demonstrate that the weighted semantics captures how varying the importance of the same factual and normative factors leads to different discretionary outcomes. The consistency across models further indicates robustness to small differences in LLM-derived weights, confirming the suitability of the approach for modeling judicial balancing. Overall, the integration of symbolic reasoning, LLM-based weighting, and formal semantics completes the pipeline, providing an interpretable and reproducible framework for analyzing discretion in law.

6. Related work

Recent work by Dik and Markovich has modeled judicial discretion through symbolic approaches that formalize its normative structure. They develop a deontic logic of judicial reasoning that characterizes discretion through duties and permissions, emphasizing how the duty of care constrains judicial freedom (Dik and Markovich 2025b), and extend it by using Answer Set Programming to encode obligations, violations, and hierarchical review between courts (Dik and Markovich 2025a). Our work, while trying to answer the "Where do weights come from?" question they also have to face, differs in both focus and methodology: rather than capturing the deontic boundaries of discretion and formalize normatively what judges ought to do, our approach descriptively reconstructs what judges do by combining symbolic methods with LLM-based weight assignment. After the early use of LLMs in hybrid symbolic-subsybolic settings (Liga 2022; Liga and Palmirani 2022; Liga and Robaldo 2023), recent work suggests that aspects of symbolic reasoning can be probed and shaped within subsymbolic models. Liga and Yu (2025) show that normative reasoning tradeoffs can be localized and causally manipulated within neural layers, suggesting an underlying mechanistic internal structure—perhaps even a norms-related *subsymbolic knowledge representation*. Complementary studies find that imposing *symbolic constraints* yields clearer explanations and supports contestability (Yu, Liga, and Markovich 2025; Freedman et al. 2025).

Formal argumentation provides a general approach to representing and reasoning with legal knowledge (Rotolo and Sartor 2023; Yu et al. 2025). Yu et al. (2024) distinguish three conceptualizations of argumentation and their roles in legal reasoning—*argumentation as inference*, *as dialogue*, and *as balancing*, and developed a meta-model (Yu and van der Torre 2025) and reasoning alignment (Rienstra, van der Torre, and Yu 2025) of formal argumentation. In this paper, we adopt WBAF (Amgoud et al. 2017; Baroni, Rago, and Toni 2019) as an instance of the balancing conceptualization, alongside related approaches such as bipolar argumentation (Yu, Markovich, and Van Der Torre 2020), preference-based frameworks (Kaci et al. 2021) and value-based models (Atkinson and Bench-Capon 2021). WBAF consists of weighted *pro* and *con* arguments, thereby fits well for modeling discretionary judicial decision making.

7. Summary and Future Work

In this paper, we have taken discretionary judicial reasoning in a real child custody dispute as a testbed to examine where numerical weights for formal models of balancing can plausibly come from. Using bipolar argumentation frameworks to capture the pros and cons articulated by three judicial levels, we prompted three LLMs (GPT-5, Claude 4.5, and DeepSeek) to assign base argument weights and evaluated them via robustness and outcome alignment. Across weighting and synthesis tasks, the LLM-generated weights were highly stable and, combined with formal argumentation semantics, reproduced the custody decisions at all three levels. This shows that LLMs can consistently recover context-sensitive weightings aligned with judicial outcomes.

A future direction is to study how LLMs assign *and adjust* the weights of arguments in discretionary judicial reasoning. Beyond assessing consistency and outcome alignment for judges, we aim to examine whether LLMs can track the meta-reasoning exhibited across the three levels of courts—specifically, how higher courts adjust the relative weight of arguments when reviewing lower-court decisions—and whether the models can justify such adjustments in ways that resemble judicial weighing. We also plan to compare LLM-generated weights with expert assessments to evaluate their plausibility and usefulness for legal reasoning to gain further insights about computationally grasping judicial discretion and the frontiers of AI.

Acknowledgments

This work was supported by the University of Luxembourg's Marie Speyer Excellence Grant *Formal Analysis of Discretionary Reasoning* (MSE-DISCREASON), and the Luxembourg National Research Fund through the project Symbolic and Explainable Regulatory AI for Finance Innovation (C24/19003061/SERAFIN).

References

Alexy, R. 2009. *A Theory of Legal Argumentation: The Theory of Rational Discourse as Theory of Legal Justification*. Oxford University Press.

Amgoud, L.; and Ben-Naim, J. 2018. Weighted bipolar argumentation graphs: Axioms and semantics. In *Twenty-Seventh International Joint Conference on Artificial Intelligence-IJCAI 2018*, 5194–5198.

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

Atkinson, K.; and Bench-Capon, T. J. 2021. Value-based Argumentation. *IfCoLog Journal of Logics and Their Applications*, 8(6): 1543–1588.

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

Cayrol, C.; and Lagasquie-Schiex, M.-C. 2009. Bipolar Abstract Argumentation Systems. *Journal of Artificial Intelligence Research*, 38: 1–36.

Dik, J.; and Markovich, R. 2024. Modeling Judicial Discretion with Nuanced Permissions. In *Legal Knowledge and Information Systems – JURIX 2024*, 48–59. IOS Press.

Dik, J.; and Markovich, R. 2025a. Judicial Discretion as Normative Reasoning – Deontic Characterization of Judicial Decision Making with Answer Set Programming. In Maranhao, J., ed., *Proceedings of the 20th International Conference on AI and Law*, 258–267. Chicago: ACM.

Dik, J.; and Markovich, R. 2025b. When Judges Go Wrong: Modeling Discretion and the Duty of Care. In van Berkel, K.; Ciabattoni, A.; and Horty, J., eds., *Deontic Logic and Normative Systems: 17th International Conference, DEON 2025*, 399–410. College Publications.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning and logic programming. *Artificial Intelligence*, 77(2): 321–357.

Dworkin, R. 2013. *Taking rights seriously*. Bloomsbury Academic.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2025. Argumentative Large Language Models for Explainable and Contestable Claim Verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14930–14939.

Hart, H. L. A. 2012. *The Concept of Law*. Oxford University Press.

Kaci, S.; van der Torre, L.; Vesic, S.; and Villata, S. 2021. Preference in Abstract Argumentation. In Gabbay, D.; Giacomin, M.; Simari, G. R.; and Thimm, M., eds., *Handbook of Formal Argumentation, Volume 2*, 211–248. College Publications.

Liga, D. 2022. Hybrid Artificial Intelligence to Extract Patterns and Rules from Argumentative and Legal Texts.

Liga, D.; and Palmirani, M. 2022. Transfer learning for deontic rule classification: The case study of the gdpr. In *Legal Knowledge and Information Systems*, 200–205. IOS Press.

Liga, D.; and Robaldo, L. 2023. Fine-tuning GPT-3 for legal rule classification. *Computer Law & Security Review*, 51: 105864.

Liga, D.; and Yu, L. 2025. Which Neurons Nudge Normative Stance? Causal Tests and Mechanistic Evidence via Contrastive Last-Token Steering. In Markovich, R.; Di Caro, L.; Rapp, A.; and Schifanella, C., eds., *Legal Knowledge and Information Systems: JURIX 2025: The Thirty-eighth Annual Conference, Turin, Italy, 9–11 December 2025*, volume 416 of *Frontiers in Artificial Intelligence and Applications*, 110–120. IOS Press. ISBN 978-1-64368-638-7.

Raz, J. 1979. *The authority of law: essays on law and morality*. New York: Oxford University Press.

Rienstra, T.; van der Torre, L.; and Yu, L. 2025. Reasoning Alignment for Agentic AI: Argumentation, Belief Revision, and Dialogue. *Journal of Applied Logics - IfCoLog Journal*, 12(6): 1683–1712.

Rotolo, A.; and Sartor, G. 2023. Argumentation and explanation in the law. *Frontiers in Artificial Intelligence*, 6: 1130559.

Shiner, R. A. 1992. *Norm and nature: the movements of legal thought*. New York: Oxford University Press.

Tucker, C. 2025. *The Weight of Reasons: A Framework for Ethics*. Oxford University Press. ISBN 978-0197786925.

Yu, L.; Liga, D.; and Markovich, R. 2025. Addressing the Right to Explanation and the Right to Challenge through Hybrid-AI: Symbolic Constraints over Large Language Models via Prompt Engineering. In *Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025)*. ACM. Forthcoming.

Yu, L.; Markovich, R.; and Van Der Torre, L. 2020. Interpretations of support among arguments. In *Legal Knowledge and Information Systems*, 194–203. IOS Press.

Yu, L.; and van der Torre, L. 2025. The A-BDI Metamodel for Human-Level AI: Argumentation as Balancing, Dialogue and Inference. In *International Conference on Logic and Argumentation*, 361–379. Springer.

Yu, L.; Van der Torre, L.; and Markovich, R. 2024. Thirteen Challenges in Formal and Computational Argumentation. *Handbook of Formal Argumentation*, 3: 931–1012.

Yu, L.; van der Torre, L.; Markovich, R.; Liao, B.; and Cai, C. 2025. DiSCo–RAD: Reasoning Alignment for Judicial Discretion. In *Logics for New-Generation AI*, 86–103. Springer.