Neurosymbolic Clinical Reasoning: From Evidence Extraction to Reason-Based Decision

Henrique Marcos[1], Foivos Ioannis Tzavellos[2] and Rohan Nanda[3,4,*]

[1]Faculty of Law, Maastricht University, The Netherlands
[2]Graduate school of Informatics, University of Amsterdam, The Netherlands
[3]Law and Tech Lab, Maastricht University, The Netherlands
[4]Department of Advanced Computing Sciences, Maastricht University, The Netherlands

Abstract

Clinical reasoning requires weighing the quality of evidence, assessing its relevance to a particular patient, and reaching a judgement that accounts for both. Doctors perform this reasoning implicitly; it is difficult to articulate and audit. Large language models can extract structured information from the biomedical literature, but their outputs lack the transparency and traceability that clinical decision-making demands. This article proposes a neurosymbolic framework in which an LLM performs the extraction and a symbolic layer performs the reasoning. The symbolic layer combines three components: Toulmin's model for organising the evidence, epistemic commitments for identifying the beliefs that bear on the decision, and a reason-based reasoning architecture for weighing those beliefs through the hierarchy of evidence and the characteristics of the patient. The framework is non-monotonic (the same evidence yields different recommendations for different patients) and auditable (every recommendation traces back through explicit reasons to source evidence). A case study on semaglutide prescription illustrates how three patient profiles produce three outcomes (prescribe, undecided, withhold) from the same evidence base, with full provenance at every stage. This position paper presents the conceptual framework; empirical validation is ongoing work.

Keywords

Neurosymbolic AI, Reason-based reasoning, Toulmin argumentation, clinical decision support, non-monotonic reasoning

1. Introduction

Consider a doctor whose patient has obesity and a history of depression. The Semaglutide Treatment Effect in People with Obesity (STEP 1) trial demonstrated that semaglutide produces mean weight reductions of 14.9 per cent in adults with obesity [1]. Pharmacovigilance analyses have reported neuropsychiatric adverse events, including suicidal ideation, though the evidence is conflicting [2, 3]. For a patient with obesity alone, the doctor prescribes. For this patient, the doctor hesitates.

The hesitation reflects a reasoning process in which the doctor weighs the quality of the evidence, assesses its relevance to this particular patient, and reaches a judgement. Doctors perform this reasoning routinely, but rarely make it explicit. It lives in clinical experience, in the informal application of the hierarchy of evidence, and in a judgement that is difficult to articulate and audit. A large language model (LLM) can extract the relevant evidence from the biomedical literature, but the extraction alone does not yield an auditable recommendation. The LLM can identify that psychiatric safety signals exist and that the STEP 1 trial reports efficacy; it cannot, in a transparent and traceable way, show why the former should outweigh the latter for this patient and not for another. For that, the reasoning must be made explicit.

Joint Workshop on Statistics and Knowledge Integration for Logic, Learning, Ethical Decisions, and LLMs (SKILLED-LLMs 2026), co-located with FLoC 2026, Lisbon, Portugal

*Corresponding author.
h.jbmarcos@maastrichtuniversity.nl (H. Marcos); f.i.tzavellos@uva.nl (F. I. Tzavellos); r.nanda@maastrichtuniversity.nl (R. Nanda)
https://henriquemarcos.com/ (H. Marcos); https://orcid.org/0009-0008-1451-5132 (F. I. Tzavellos); https://www.maastrichtuniversity.nl/nl/r-nanda (R. Nanda)
0000-0003-0259-5530 (H. Marcos); 0009-0008-1451-5132 (F. I. Tzavellos); 0000-0001-7124-4799 (R. Nanda)

This article proposes a neurosymbolic framework that separates extraction from reasoning. The LLM extracts structured evidence from PubMed abstracts; a symbolic layer reasons over that evidence using a formal scaffolding that mirrors what doctors already do. The scaffolding has three components. Toulmin's model of argumentation [4] provides the vocabulary for organising the evidence. *Epistemic commitments* extract from the Toulmin structure the beliefs that bear on the clinical decision. *Reason-based reasoning* (RBR), grounded in the tradition of reason-based logic [5], provides the formal structure for weighing those beliefs using the hierarchy of evidence, adjusted for the characteristics of the patient. Because the reasoning is performed by the symbolic layer, every recommendation is auditable: traceable through explicit reasons, at multiple levels, back to the source evidence.

Fox and Modgil [6] identified the gap between evidence and decision in Toulmin's model and proposed extending it toward decision-making through the qualitative aggregation of arguments. This article develops that programme by integrating Toulmin with RBR. Three contributions follow: the bridge from Toulmin to RBR through epistemic commitments; a neurosymbolic architecture that ensures the reasoning remains transparent and inspectable; and a demonstration of non-monotonicity (same evidence, different patient, different recommendation) with full provenance.

2. The Proposed Framework

2.1. Structuring the Evidence

Reading the semaglutide literature, one encounters claims supported by trial data, qualified by methodological limitations, and challenged by safety concerns. Toulmin's model [4] provides a vocabulary for this organisation. The claim is what the studies say (semaglutide produces significant weight loss); the data is what the trials found; the warrant is why the data supports the claim (consistent findings across multiple RCTs). Every claim, however, has boundaries: the qualifier marks where the evidence is limited (studies primarily involve adults without diabetes; weight regain occurs after treatment discontinuation), and the rebuttal marks what might undermine it (gastrointestinal adverse events; conflicting neuropsychiatric safety data). The Toulmin model tells us what the evidence says; the question of what to do requires a further step.

2.2. Extracting Beliefs

The next step is to extract from the Toulmin structure the beliefs that bear on the decision to prescribe: the drug works, stopping leads to regain, there are safety concerns. These are *epistemic commitments*: justified beliefs, grounded in the evidence, that function as the considerations to be weighed. The justification traces back to the Toulmin structure; the warrant says why the belief is justified, the backing says why the warrant holds. To the extent that the states of affairs these beliefs describe obtain, a correspondence theory of truth yields a compact result: a justified true belief corresponds to a fact, that is, a state of affairs that obtains [5]. A fact that pleads for or against a conclusion is a reason. The same object is therefore simultaneously a belief (from the clinician's perspective), a fact (from the ontological perspective), and a reason (from the perspective of the decision); 'epistemic commitment' names the object at the bridge between evidence and decision.

We propose a schema of epistemic commitment categories that the efficacy evidence of drugs typically yields: efficacy (EC-EFF: the drug produces the intended therapeutic effect), dependency (EC-DEP: stopping treatment leads to an adverse outcome), population scope (EC-POP: the evidence covers a specific population), uncertainty (EC-UNC: long-term effects remain unknown), and risk (EC-RISK, further distinguished by type: gastrointestinal, psychiatric). These categories are organised by the hierarchy of evidence that the medical profession employs [7], adapted here to operate by evidential role: Tier 1 comprises primary endpoint evidence from RCTs (what the trials were designed to measure); Tier 2 comprises secondary safety observations from the same trials (adverse events collected alongside the primary endpoint); Tier 3 comprises post-marketing pharmacovigilance data; and Tier 4 comprises qualifiers and absence of evidence, which carry no independent polarity (Table 1). Note that tiering reflects evidential role in the decision, not study design alone. Dependency data from RCTs are classified as Tier 4 because they qualify the efficacy claim rather than independently supporting prescription or withholding; population-scope and long-term-uncertainty commitments are Tier 4 because they flag limits of the evidence without pleading either way.

2.3. Weighing the Reasons

The clinical decision involves two incompatible conclusions: D_{presc} (prescribing is justified) and D_{with} (withholding is justified). Each epistemic commitment pleads for one and, by the incompatibility, against the other. The question is which side the balance of reasons supports, and the framework within which this question receives a precise formulation is reason-based reasoning (RBR).

RBR operates with a single ontological category: *states of affairs*. Those that obtain are facts. Some facts obtain independently of the balance of reasons (that semaglutide was tested in the STEP 1 trial is such a fact). Others obtain only if the reasons for their obtaining outweigh the reasons against; that prescribing is justified for this patient is a state of affairs of this second kind. These are *reason-dependent* states of affairs. The full RBR framework [8, 5] comprises ten definitions covering the application of norms, their conflicts, and their resolution through reasons at multiple levels. The clinical case requires only the fragment governing reason-dependent states of affairs, for which five definitions suffice.

Notation. F denotes a state of affairs; those that obtain are facts. D denotes a reason-dependent state of affairs. $pro(D)$ and $con(D)$ denote the sets of reasons that plead for and against D obtaining. The symbol $\succ$ denotes the outweighing relation. K denotes the set of operative constraints and W_K the set of admissible worlds satisfying K [9].

Definition 1 (Reason). A state of affairs F is a reason if and only if F obtains (is a fact) and there exists a reason-dependent state of affairs D such that F pleads for or against D obtaining.

$$Reason(F) \leftrightarrow Obtains(F) \ \wedge \ \exists D \ : \ F \in pro(D) \vee F \in con(D)$$

Efficacy evidence, safety data, and patient characteristics all bear on whether to prescribe. Each is a reason in the sense of Definition 1.

Whether a fact 'pleads for or against' a conclusion is case-relative. Gastrointestinal adverse events are a fact that obtains regardless of who the patient is, but that fact only pleads against prescribing for a patient with gastrointestinal vulnerability. For a patient without such vulnerability, the fact is noted but does not enter the weighing; it is a fact, but it is not a reason for this patient. The patient's characteristics determine which facts function as reasons.

Definition 2 (Obtaining). A reason-dependent state of affairs D obtains if and only if the set of pro reasons for D outweighs the set of con reasons.

$$Obtains(D) \leftrightarrow pro(D) \succ con(D)$$

Prescribing is justified if the reasons for it outweigh the reasons against. Definition 2 deliberately leaves $\succ$ unspecified; the framework provides structure, the domain provides substance. In this case the domain supplies that substance through the hierarchy of evidence: F_{hier} orders reasons by evidential tier, and F_{caut} gives priority to existentially severe risks. Where neither resolves a contest, the framework returns undecided rather than imposing a verdict.

Definition 3 (Default weighing). For any sets of reasons S and S', if S is empty and S' is non-empty, then S' outweighs S.

$$S = \emptyset \ \wedge \ S' \neq \emptyset \ \rightarrow \ S' \succ S$$

When there are reasons on one side and nothing on the other, the decision follows. Definition 3 establishes the baseline; the interesting cases arise when both sides have reasons.

The domain of admissible outcomes is bounded by *constraints*: conditions that exclude certain combinations of states of affairs and cannot themselves be outweighed. One cannot simultaneously

Table 1
Epistemic commitments for semaglutide prescription

EC	Category	Content	Tier	PubMed PMIDs
EC-EFF	Efficacy	14.9% weight loss	1	33567185
EC-DEP	Dependency	Stopping → weight regain	4	35441470
EC-RISK-GI	Gastrointestinal (GI) risk	May cause discontinuation	2	33567185, 36578889
EC-RISK-PSYCH	Psychiatric risk	Contested signals	3	40523410, 38182782
EC-POP	Population scope	Primarily adults without diabetes	4	33567185
EC-UNC	Uncertainty	Long-term effects remain uncertain	4	40904035

prescribe and withhold; this incompatibility is a constraint. The set of admissible worlds W_K comprises those in which all constraints in K are satisfied [9]. Constraints narrow the space of what is possible; reasons operate within what remains.

Definition 4 (Compatibility). Two states of affairs are compatible if and only if they can obtain simultaneously within at least one admissible world.

$$Compatible(F, F') \leftrightarrow \exists w \in W_K : F \in w \wedge F' \in w$$

D_{presc} and D_{with} are incompatible: no admissible world contains both.

Definition 5 (Cross-conflict). If F and F' are incompatible, then any pro reason for F is a con reason for F', and vice versa.

$$\neg Compatible(F, F') \rightarrow pro(F) \subseteq con(F') \wedge pro(F') \subseteq con(F)$$

The efficacy evidence that supports prescribing is, by implication, a consideration against withholding. Definition 5 generalises the attack relation of Dung's abstract argumentation frameworks [10] to the setting of reason-dependent states of affairs: the conflict arises structurally from the incompatibility of the conclusions, not from a stipulated relation between the reasons.

Where both sides have reasons and the balance cannot be resolved at the first order, the question of which side outweighs is itself a reason-dependent state of affairs, governed by Definition 2 at the next order. A second-order reason bears on the weight of first-order reasons; a third-order reason bears on the weight of second-order reasons. The recursion terminates because the pool of available reasons is finite and subject to a non-recycling condition: no reason that appears at order n for a given target may appear at any order $m > n$ for the same target.

Two facts govern the higher orders in the clinical case. F_{hier} is the fact that the medical profession ranks primary endpoint evidence above pharmacovigilance data; this fact functions as a second-order reason because it bears on the weight of first-order reasons. F_{caut} is the fact that, when a patient is specifically vulnerable to a reported harm, the asymmetric severity of error favours caution; this fact functions as a third-order reason when the potential harm is existentially severe.

3. Case Study: Semaglutide

PubMed abstracts covering semaglutide efficacy, safety, and long-term outcomes were retrieved and used to construct an illustrative evidence base. The case is illustrative; the scenarios demonstrate the framework's reasoning structure and should not be interpreted as clinical recommendations. From this evidence base, we identified the epistemic commitments summarised in Table 1 through LLM-assisted analysis and manual verification against the source literature. D_{presc} and D_{with} are incompatible; by Definition 5, the pro reasons for one are con reasons for the other. Three patient profiles illustrate how the framework resolves (or declines to resolve) the resulting contest.

Scenario A: Obesity only. The patient has obesity and no pre-existing conditions. Efficacy evidence (EC-EFF) pleads for D_{presc}. Risk commitments require matching patient vulnerabilities, and this patient

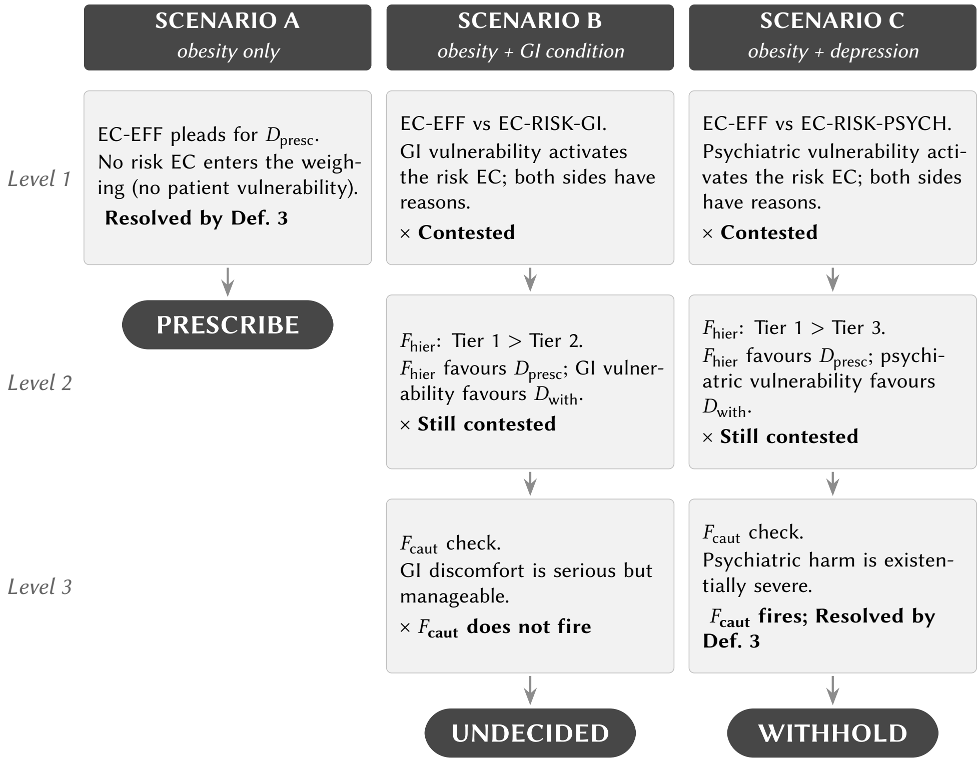

Figure 1: Multi-level resolution across the three scenarios. Scenario A resolves at Level 1; Scenarios B and C are contested through Level 2 and diverge at Level 3, where F_{caut} fires only when the potential harm is existentially severe. The evidence base is identical across patients; the case facts determine which facts function as reasons.

has none; no risk EC enters the weighing. EC-DEP, EC-POP, and EC-UNC are qualifiers that carry no independent polarity. At Level 1, $pro(D_{presc})$ is non-empty and $con(D_{presc})$ is empty. By Definition 3, D_{presc} obtains. Recommendation: prescribe.

Scenario B: Obesity with GI condition. The patient has obesity and a pre-existing gastrointestinal condition. EC-RISK-GI now pleads against D_{presc}, because the patient's vulnerability makes gastrointestinal adverse events specifically relevant. At Level 1, both sides have reasons (EC-EFF for prescribing, EC-RISK-GI against) and the contest is unresolved. At Level 2, F_{hier} enters as a second-order reason: EC-EFF is Tier 1 evidence (primary endpoint) and EC-RISK-GI is Tier 2 (secondary safety observation), so the hierarchy of evidence favours D_{presc}. The patient's gastrointestinal vulnerability, however, enters as a competing second-order reason favouring D_{with}: for this patient, the secondary safety observation is specifically relevant. Both F_{hier} and the patient's gastrointestinal vulnerability are present, and the second-order contest is unresolved. At Level 3, gastrointestinal discomfort is serious but manageable, so F_{caut} does not fire. No third-order reason resolves the contest. Recommendation: undecided. The framework presents both sides for the clinician.

Scenario C: Obesity with depression. The patient has obesity and a history of depression. EC-RISK-PSYCH now pleads against D_{presc}, because the patient's psychiatric vulnerability makes the contested neuropsychiatric signals specifically relevant. The reasoning follows the same structure as Scenario B: Level 1 is contested; at Level 2, F_{hier} favours D_{presc} (EC-EFF is Tier 1, EC-RISK-PSYCH is Tier 3), while the patient's psychiatric vulnerability favours D_{with}. The difference emerges at Level 3. A psychiatric adverse event in an already vulnerable patient is existentially severe, so F_{caut} fires, supporting D_{with}. F_{caut} is unopposed at the third order; by Definition 3, D_{with} obtains at the third order, which in turn resolves the second-order contest in favour of D_{with}. Recommendation: withhold. Figure 1 summarises the level-by-level resolution for all three scenarios.

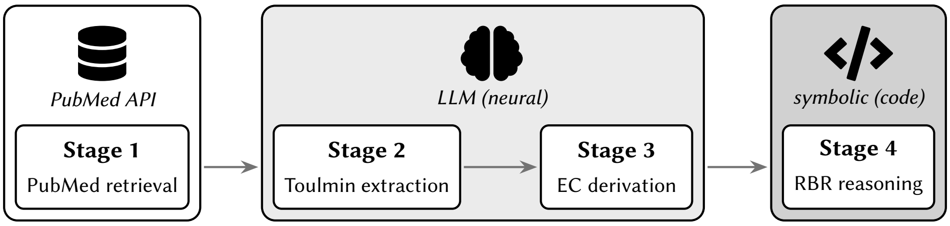

Figure 2: The four-stage neurosymbolic pipeline. Each container groups stages by implementation: the two neural stages share one LLM box, the symbolic stage its own. Stage 4 implements Definitions 1–5.

4. Computational Realisation

This section describes the proposed pipeline architecture; implementation and empirical validation are ongoing work. The framework is designed as a four-stage pipeline. PubMed abstracts are retrieved via API queries structured around Toulmin categories; an LLM extracts a consolidated Toulmin structure with explicit citations and then derives epistemic commitments through prompting guided by the Epistemic Commitment (EC) schema of Section 2.2, outputting structured JSON with EC type, polarity, contested flag, and source PMIDs. Manual verification validates the extracted commitments against the source literature. The final stage is a symbolic layer that performs multi-level RBR. This stage involves no LLM; it is deterministic code implementing Definitions 1–5. Given epistemic commitments and case facts, it produces an outcome (prescribe, undecided, or withhold) together with the level of resolution and a full provenance trace. This paper presents the conceptual architecture; implementation details, including model selection, prompt design, and empirical validation, are the subject of ongoing work. The division of labour is deliberate: the LLM extracts, the symbolic layer reasons. Because the reasoning structure is explicit, inspectable, and independent of the neural model, this separation addresses the transparency requirements that medical AI demands [11] while leveraging the extraction capabilities that LLMs provide. Figure 2 illustrates this architecture.

5. Discussion and Conclusion

The three scenarios yield prescribe, undecided, and withhold from the same evidence base. No new evidence was introduced between them: the same Toulmin structure, the same epistemic commitments, the same first-order reasons. What changed was a single case fact, which created a contest that required higher-order resolution. This reveals the non-monotonic character of the framework: F_{hier} was defeated in Scenario C, but what was defeated was the conclusion F_{hier} supported, not F_{hier} itself. F_{hier} remains a fact and a reason; under different case facts (Scenario A) it prevails. The undecided outcome in Scenario B is equally significant; the framework recognises when the evidence and the patient's characteristics do not provide sufficient grounds for resolution, and it says so.

Every element of the conclusion is traceable. D_{with} in Scenario C rests on a third-order resolution: at Level 1, EC-RISK-PSYCH pleads for withholding; at Level 2, F_{hier} and the patient's psychiatric vulnerability contest; at Level 3, F_{caut} resolves. A clinician who disagrees can, in principle, locate the point of disagreement, whether the existence of the safety concern, the relevance of the patient's history, or the applicability of the precautionary principle.

The formal apparatus connects to several established frameworks in the argumentation literature. Definition 5 generalises the attack relation of Dung's abstract argumentation frameworks [10] to the setting of reason-dependent states of affairs. Bipolar argumentation frameworks [12] distinguish support and attack, and Abstract Dialectical Frameworks [13] allow arbitrary acceptance conditions; RBR provides a specific semantics in which contested weighing at Level n is resolved by reasons at Level n+1. A survey of the argumentation literature [14] documents that existing approaches have not sufficiently addressed the logical encoding of weighted argumentation with multi-level resolution.

Limitations remain. The EC typology may not capture all clinically relevant beliefs; comparative efficacy and considerations of cost are absent. The case study is illustrative, and the hierarchy of evidence is a working proposal. The framework's reliability depends on the quality of LLM extraction; hallucinated or misclassified evidence would propagate into the symbolic layer, and mitigation strategies require further investigation. Any decision-support system carries the risk of automation bias; the framework mitigates this by exposing every reason and its provenance, and by returning undecided rather than a verdict when reasons do not resolve, leaving the determination with the clinician. How practitioners use such traces is an empirical question for future work. Future work includes a pipeline integrating LLM-based extraction with an implementation of symbolic multi-level RBR, annotation studies for validation, and connection to ASPIC+ [15] for structured argumentation.

Declaration on Generative AI

During preparation of this work, the authors used Claude for sentence polishing and rephrasing. After using it, the authors reviewed and edited the content and take full responsibility for the paper's content.

References

[1] J. P. H. Wilding, R. L. Batterham, S. Calanna, M. Davies, L. F. Van Gaal, I. Lingvay, B. M. McGowan, J. Rosenstock, M. T. D. Tran, T. A. Wadden, S. Wharton, K. Yokote, N. Zeuthen, R. F. Kushner, Once-weekly Semaglutide in Adults with Overweight or Obesity, New England Journal of Medicine 384 (2021) 989–1002. doi:10.1056/NEJMoa2032183.

[2] M. Wang, Z. Yang, M. Yan, S. Liu, S. Xiao, Depression and suicide/self-injury signals for weight loss medications: A disproportionality analysis of semaglutide, liraglutide, and tirzepatide in FAERS database, Journal of Affective Disorders 389 (2025) 119670. doi:10.1016/j.jad.2025.119670.

[3] W. Wang, N. D. Volkow, N. A. Berger, P. B. Davis, D. C. Kaelber, R. Xu, Association of semaglutide with risk of suicidal ideation in a real-world cohort, Nature Medicine 30 (2024) 168–176. doi:10.1038/s41591-023-02672-2.

[4] S. E. Toulmin, The Uses of Argument, 2 ed., Cambridge University Press, 2003. doi:10.1017/CBO9780511840005.

[5] J. Hage, Reasoning with Rules: An Essay on Legal Reasoning and its Underlying Logic, Springer, Dordrecht, 1997. doi:https://doi.org/10.1007/978-94-015-8873-7.

[6] J. Fox, S. Modgil, From Arguments to Decisions: Extending the Toulmin View, in: D. Hitchcock, B. Verheij (Eds.), Arguing on the Toulmin Model, Springer Netherlands, Dordrecht, 2006, pp. 273–287. doi:10.1007/978-1-4020-4938-5_18.

[7] P. B. Burns, R. J. Rohrich, K. C. Chung, The levels of evidence and their role in evidence-based medicine, Plastic and Reconstructive Surgery 128 (2011) 305–310. doi:10.1097/PRS.0b013e318219c171, pMID: 21701348.

[8] H. Marcos, Lex Specialis as a Reason-Giving Norm: Balancing Norm Specificity and Individual Rights in Times of Crisis, International Community Law Review 27 (2025). doi:10.1163/18719732-bja10138.

[9] S. A. Kripke, Semantical considerations on modal logic, Acta Philosophica Fennica 16 (1963) 83–94.

[10] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games, Artificial Intelligence 77 (1995) 321–357. doi:10.1016/0004-3702(94)00041-X.

[11] C. Kim, S. U. Gadgil, S.-I. Lee, Transparency of medical artificial intelligence systems, Nature Reviews Bioengineering 4 (2026) 11–29.

[12] C. Cayrol, M.-C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2005), volume 3571 of *LNCS*, Springer, 2005, pp. 378–389. doi:10.1007/11518655_33.

[13] G. Brewka, S. Woltran, Abstract dialectical frameworks, in: Proceedings of the Twelfth International Conference on Principles of Knowledge Representation and Reasoning (KR 2010), AAAI Press, 2010, pp. 102–111.

[14] P. Besnard, C. Cayrol, M.-C. Lagasquie-Schiex, Logical theories and abstract argumentation: A survey of existing works, Argument & Computation 11 (2020) 41–102. doi:10.3233/AAC-190476.
[15] H. Prakken, An abstract framework for argumentation with structured arguments, Argument & Computation 1 (2010) 93–124. doi:10.1080/19462160903564592.