A Framework for Representing and Quantifying Contestation of Claims for AI Debate Evaluation

Neal Shah[1], Rishi Mukherjee[1,2]

[1]DebateArena
[2]Harvard Graduate School of Education

Abstract

This paper introduces the first component of a framework that formally represents debates for AI evaluation using claims and prove/disprove relations between claims. We focus on contestation, or a claim's standing, in this framework. AI evaluates debates by assigning recursively updated weights to claims using the Cesàro average to resolve cyclic argumentation problems. A prototype of this framework has been used to evaluate 9,500 debates on the online debate platform DebateArena.

Keywords

computational argumentation, argumentation semantics, cyclical argumentation, contestation, debate evaluation, AI judge, artificial intelligence, debate

1. Introduction

Evaluating who won a debate requires identifying the claims in a transcript of arguments, and interpreting the relations between claims. Future research will explore all the types of relations between arguments. We focus on the "contestation" relation i.e. whether a claim is disputed. We represent the contestation of arguments in a debate as a graph of prove and disprove relations and assign each claim a weight from 0 to 1 denoting its standing. We can evaluate debates by recursively updating weights according to the weight of other claims that prove or disprove a claim.

Our framework is claim-atomic, meaning claims are the smallest argumentative unit. Because claims contest one another recursively the weights are mutually dependent and can cause "cycles" e.g. Claim A disproves Claim B which disproves Claim A, and we resolve this issue by defining finalized weights as the Cesàro average of the weights over time e.g. 0, 1, 0, 1 becomes a weight of .5, which is one of our central contributions. The prototype for our framework runs on the debate education platform DebateArena, where AI adjudication lets several debates proceed at once in a class that one teacher could not otherwise judge.

2. Related Work

We are inspired heavily by Dung's [1] way of modeling argumentation as arguments with relations in between, deliberately abstracting away the internal structure. Our atomic unit is instead claims, but we define the relations in between in a bipolar manner (even if just for contestation) like Cayrol & Lagasquie-Schiex [2]—our two relations between claims being "prove" and "disprove". Like many others before us, we represent contestation with recursively updated weights, drawing on the gradual semantics tradition [3]. Our specific framework results in cycles because of processes like self-disproof and mutually disproving claims, which are present at high levels of debate. Cyclical argumentation has already been dealt with by Cayrol & Lagasquie-Schiex [7], where they have discussed golden-ratio strengths for arguments in cycles. That case converges, so a limit suffices there; our Remark 1 recovers the same phenomenon, and the Cesàro average agrees with their limit. Recently, closely related work

CMNA'26: The 26th International Workshop on Computational Models of Natural Argument, September 14–15, 2026, Barcelona, Spain

*Corresponding author.

has been done by Gabbay & Rodrigues [6]. Specifically, they also propagate weights between 0 and 1 and evaluate them until they settle, though their weights are instantiated outside the framework, and, like Potyka [4], their update is continuous and damping, which allows cycles to reach an equilibrium. We cannot adopt their construction because their convergence argument requires continuity, which our convention $0/0 := 1$ violates. Our update is expressible as an ADF [5] acceptance condition over $[0, 1]$, though the final assignment is not one of the fixed-point semantics ADFs are typically given.

3. Framework

Definition 1 (Claims and contestation graph)**.** *Let C be a finite set of atomic claims; all content in the debate is a claim and claims are the framework's sole primitives. There are three binary contestation relations between claims: y proves x, y disproves x, or y has no contestation effect on x. We represent these as a directed signed graph $G = (C, E)$ with $E \subseteq C \times C \times \{+, -\}$, absence of an edge denoting no contestation effect, and for each x collect $P(x) = \{y : y \xrightarrow{+} x\}$ and $D(x) = \{y : y \xrightarrow{-} x\}$.*

Definition 2 (Contestation weight)**.** *A contestation weight is a map $w : C \to [0, 1]$,*

$$w(x) = \frac{\sum_{y \in P(x)} w(y)}{\sum_{y \in P(x)} w(y) + \sum_{y \in D(x)} w(y)}, \qquad 0/0 := 1.$$

An only-supported claim has weight 1, an only-opposed one 0, and an unchallenged claim has weight $0/0 := 1$. The definition is recursive (e.g. a challenge from a disproven claim carries no weight).

Definition 3 (Contestation assignment)**.** *Let $\Phi : [0, 1]^C \to [0, 1]^C$ apply the formula of Definition 2 to every claim at once. The contestation assignment is*

$$\bar{w} \;=\; \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \Phi^n(\mathbf{1}).$$

Proposition 1. *If G is acyclic, a unique contestation assignment exists and is computed by a single topological sweep from the sources, with no iteration or averaging; uniqueness follows by induction along a topological order.*

Remark 1. Cycles need not converge, and a trajectory may fail to repeat even when it does. Two claims which mutually disprove one another, run from full standing, oscillate between 1 and 0, so each receives standing $\frac{1}{2}$. A claim s proving a claim t which disproves itself updates t by $w \leftarrow 1/(1 + w)$, producing ratios of consecutive Fibonacci numbers converging to $1/\varphi \approx 0.618$ but never repeating. The Cesàro average handles both: it is regular, agreeing with the limit when one exists and extending it when none does.

Remark 2. We conjecture that the contestation assignments are Cesàro-summable for every finite signed graph. As shown above, this is true in the acyclic case, and an exhaustive check over all 19,683 signed graphs of three nodes, including self-loops, found no exception.

Worked example. Take A "global warming is real", proven by B "IPCC reports confirm it" and disproven by C "there is an ice age now", itself disproven by D "satellites show Arctic ice has melted" and D by E "ice is invisible on satellites". The graph is acyclic, so Proposition 1 applies: B and E are unanswered and take weight 1; $w(D) = 0/(0 + w(E)) = 0$, so D exerts no force on C and $w(C) = 1$ by the $0/0$ convention; and $w(A) = w(B)/(w(B) + w(C)) = 1/2$, the claim being equally proven and disproven.

Future work. We are pursuing a proof of our conjecture in Remark 2, assembling an empirical comparison against human adjudications on DebateArena, and extending the relation set to mitigation and priority relations.

Declaration on Generative AI

For the paper's preparation, the authors used generative AI to check grammar and the formal work. The authors then reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] C. Cayrol and M.-C. Lagasquie-Schiex, Bipolar abstract argumentation systems, in: *Argumentation in Artificial Intelligence*, Springer, 2009.

[3] L. Amgoud and J. Ben-Naim, Ranking-based semantics for argumentation frameworks, in: *Proceedings of SUM 2013*, Springer, 2013, pp. 134–147.

[4] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: *Proceedings of KR 2018*, AAAI Press, 2018, pp. 148–157.

[5] G. Brewka, S. Ellmauthaler, H. Strass, J. Wallner, and S. Woltran, Abstract dialectical frameworks revisited, in: *Proceedings of IJCAI 2013*, 2013, pp. 803–809.

[6] D. M. Gabbay and O. Rodrigues, Equilibrium states in numerical argumentation networks, *Logica Universalis* 9 (2015) 411–473.

[7] C. Cayrol and M.-C. Lagasquie-Schiex, Graduality in argumentation, *Journal of Artificial Intelligence Research* 23 (2005) 245–297.