Mathematical Formalization of Persuasive Arguments Theory: Explaining Group Polarization and Choice Shifts

David Rieger

Author Note

Student ID: 12558391

Course: P 19: Empirisch-psychologisches Praktikum

Date: 22.02.2026

Abstract

[Abstract zur mathematischen Formalisierung der Persuasive Arguments Theory hier einfügen]

Keywords: Persuasive Arguments Theory, Group Polarization, Choice Shifts, Mathematische Modellierung

Mathematical Formalization of Persuasive Arguments Theory: Explaining Group Polarization and Choice Shifts

Introduction

Objective of the Formalization

The primary objective of the following work is to translate the verbal formulations of Persuasive Arguments Theory (PAT) into a formalized, mathematical, and computational model to explicitly explain the phenomenon of group polarization This formalization adheres to the theoretical specification framework proposed by Van Lissa et al. (2026), ensuring that the resulting model aligns with FAIR principles — making the theory Findable, Accessible, Interoperable, and Reusable.

Introduction to Persuasive Arguments Theory

Persuasive Arguments Theory (PAT), originally formulated by Burnstein and Vinokur (1977), is a cognitive model designed to explain attitude shifts through the mechanisms of information processing. As highlighted by Isenberg (1986), PAT posits that an individual's position on an issue is determined by the number and persuasiveness of arguments they can recall from memory when formulating their own opinion. Isenberg (1986) strictly contrasts this cognitive focus with social comparison theories, which argue that shifts occur primarily due to individuals' desires to conform to or exceed group norms to maintain social desirability. Crucially, the meta-analysis by Isenberg (1986) concludes that PAT provides a significantly more robust empirical explanation for the magnitude and direction of group polarization than social comparison. Currently, there are very few papers that mathematically formalize PAT. While a recent paper does so using an agent-based model, it focuses specifically on the boundary conditions for quasiconsensus and bipolarization (Pedraza et al., 2025). Explicit computational formalizations targeting

the specific phenomenon of group polarization, which the theory was originally created to explain, remain absent.

Phenomenological Target: Group Polarization

The phenomenological target of this formalization is group polarization. Group polarization is defined as the occurrence where an initial tendency of individual group members toward a given direction is enhanced following group discussion (Isenberg, 1986). The core mechanism dictates that on decisions where group members hold a moderate average proclivity in a specific direction, subsequent group discussion results in a more extreme average proclivity in that identical direction.

A prototypical primary study demonstrating this effect was conducted by Moscovici and Zavalloni (1969). In their experimental design, participants initially recorded their private opinions and judgments on various issues. They subsequently discussed these issues in groups of four to reach a consensus, which was followed by a final private recording of their opinions. The results consistently demonstrated that initial attitudes were significantly intensified post-discussion.

The robustness of group polarization is extensively documented. Isenberg (1986) critical review and meta-analysis, alongside foundational primary literature (Moscovici & Zavalloni, 1969), provides strong empirical evidence for the phenomenon's magnitude and replicability. Assessing generalizability through the Units, Treatments, Outcomes, and Settings (UTOS) framework (Cronbach & Shapiro, 1982), group polarization demonstrates broad applicability based on the data aggregated by Isenberg (1986). It has been reliably observed across diverse demographic units (e.g., students, simulated juries), various experimental treatments (e.g., differing discussion formats), multiple operationalized outcomes (e.g., choice dilemmas, mock-trial judgments), and diverse laboratory settings. The phenomenon can be considered robust because the empirical evidence is consistently

strong, and the effect generalizes reliably across the constituent dimensions of the UTOS framework present in the meta-analytic literature.

Methods

All computational code, simulation data, and supplementary materials required to reproduce this formalization are openly accessible in the project repository at https://github.com/David-Rieger/Persuasive-Arguments-Theory.

Core Constructs of Persuasive Arguments Theory

To formalize Persuasive Arguments Theory (PAT), we systematically map its verbal formulations into a Visual Argument Structure Tool (VAST) display, a graphical framework designed to represent theoretical assumptions explicitly (Leising et al., 2023). As illustrated in Figures 1, 2, and 3, the VAST models categorize the theory into three distinct levels of resolution: the phenomenological level, the theoretical (verbal) level, and the argument level. For a complete mapping of all extracted concepts, original quotes, and their corresponding IDs, please refer to the Construct Source Table in Appendix A.

At the macro-level (group level), PAT revolves around the initial tendency of the group with a moderate average proclivity ($\overline{T}^{pre}$; ID 17), the resulting tendency of the group after group discussion ($\overline{T}^{post}$; ID 18), and the resulting phenomenon of Group Polarization (GP; ID 1). Constructs $\overline{T}^{pre}$ and $\overline{T}^{post}$ were developed independently for structural completeness to quantify aggregate group states, as the original literature explicitly operationalizes tendencies strictly at the individual level.

At the micro-level (individual level), these macro-constructs are derived from the individual's initial tendency (T_i^{pre}; ID 2) and their resulting tendency post-discussion (T_i^{post}; ID 4). The difference between these two temporal states defines the individual choice shift (CS; ID 11), which is triggered by the group discussion (GD; ID 3).

Finally, at the argument level, the cognitive foundation of these tendencies is quantified. The central constructs include the initial argument set (IA_i; ID 5) and the additional persuasive arguments encountered during discussion (AA_i; ID 9). These sets are characterized by the absolute number of arguments (n_i^{IA}, ID 6; n_i^{AA}, ID 10) and the specific properties of each argument: its directional valence (d_j; ID 7) and its persuasiveness (p_j; ID 7, 13), with the latter being a function of its perceived validity (val_j; ID 14) and perceived novelty (nov_j; ID 15).

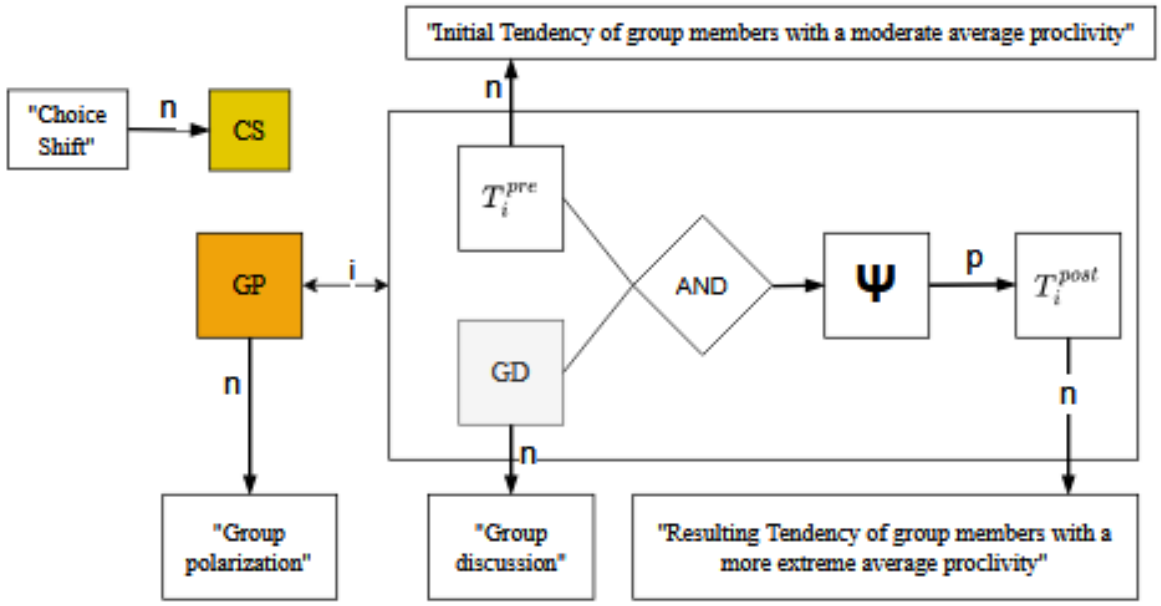
Theoretical Mechanisms and Phenomenon Emergence

The VAST models visually deconstruct how PAT explains the emergence of group polarization from underlying cognitive processes. Parenthetical IDs reference the Construct Source Table (Appendix A), specifically denoting theoretical relationships rather than isolated constructs as in the previous section.

In the first panel, shown in Figure 1, the macro-phenomenon of Group Polarization (GP) is defined as the aggregate consequence of individual-level choice shifts (CS). When individuals with a moderate initial tendency (T_i^{pre}) engage in group discussion (GD), psychological processing occurs (represented by the Ψ-box), resulting in an updated, more extreme tendency (T_i^{post}).

Figure 1

VAST Display Illustrating the Phenomenon of Group Polarization and Individual Choice Shift

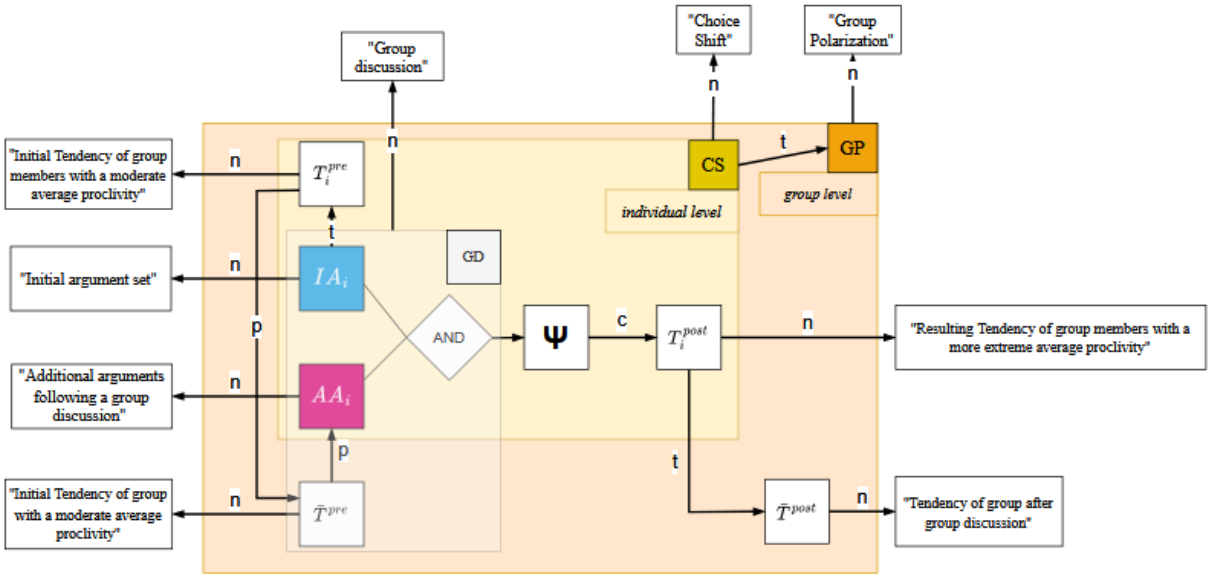


Note. Mapping of the transition from individual choice shifts to aggregate group polarization.

The second panel, shown in Figure 2, elaborates on the inputs of the Ψ-box. It illustrates that the transition from T_i^{pre} to T_i^{post} is causally driven by the cognitive integration of the individual's initial argument set (IA_i) and the novel, additional arguments (AA_i) retrieved during the discussion (ID 8, ID 12). Crucially, we maintain the Ψ-box as a "closed" black box in this diagram. Because all measurable variables (i.e., prior tendencies, argument counts, and resulting tendencies) are already explicitly modeled as external inputs and outputs, opening the Ψ-box to duplicate these variables internally would have caused severe visual clutter and redundancy without adding structural clarity.

Figure 2

VAST Display Illustrating the Theoretical Level of Persuasive Arguments Theory

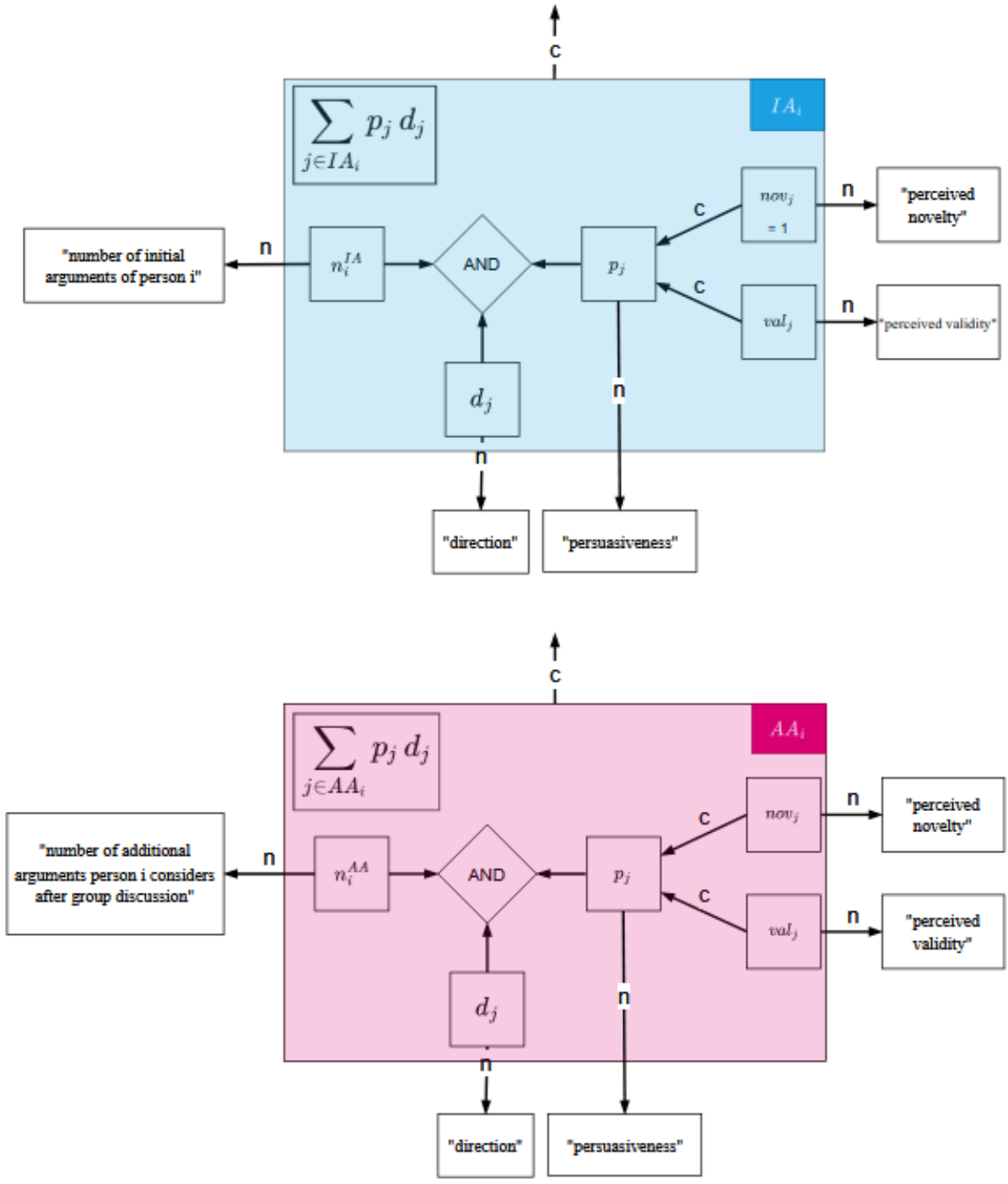


Note. Conceptual framework for the cognitive integration of initial and additional argument sets.

The third panel, shown in Figure 3, specifies the causal determinants of an argument's persuasiveness, explaining *why* an argument successfully induces a choice shift. According to PAT, the persuasiveness of any given argument (p_j) is a direct function (ID 16) of its perceived validity (val_j) and its perceived novelty (nov_j). Through the mathematical summation of the persuasiveness (p_j) and direction (d_j) of all considered arguments (IA_i and AA_i), the individual's cognitive architecture generates a choice shift. When these individual shifts align directionally, they aggregate at the group level to produce the robust phenomenon of group polarization.

Figure 3

VAST Display Illustrating the Argument Level and Determinants of Persuasiveness



Note. Structural breakdown of variables determining argument persuasiveness.

Mathematical Formalization and Model Architecture

The computational translation of Persuasive Arguments Theory relies on a specific set of quantified variables and mathematically defined relationships. For a comprehensive overview, the Variable Table and the Relationships Table are provided in Appendix A.

At the macro-level, the initial tendency of the group ($\overline{T}^{pre}$) and the resulting tendency of the group after discussion ($\overline{T}^{post}$) are formalized on a continuous scale ranging from -1 to +1, where -1 indicates an extremely negative opinion and +1 indicates an extremely affirmative opinion on a given topic. Similarly, at the micro-level, the initial tendency of individual group members (T_i^{pre}) and their resulting tendency post-discussion (T_i^{post}) operate on an identical continuous scale from -1 to +1, utilizing the same anchoring endpoints.

The cognitive processing of information is quantified through argument-level variables. The volume of information is represented by the absolute number of initial arguments (n_i^{IA}) and additional arguments (n_i^{AA}), both defined as discrete integers ($0, 1, 2, \cdots \in \mathbb{N}_0$). Each individual argument is characterized by its directional valence (d_j), which is captured on a dichotomous scale restricted to -1 (representing a contra argument) and +1 (representing a pro argument). Furthermore, the persuasiveness of each argument (p_j) is defined on a continuous scale from 0 to 1, anchored at 0 for an argument that is not persuasive at all and 1 for an argument that is maximally convincing. To streamline the computational architecture, the theoretically proposed sub-components of perceived novelty and perceived validity were deliberately excluded from the direct parameter space. Their theoretical influence is instead exclusively aggregated into the singular persuasiveness factor (p_j).

Baseline Mathematical Relationships. The core causal mechanisms of the model are formalized through the following baseline mathematical relationships.

The individual's initial argument set (IA_i) and additional argument set (AA_i) are

defined as arrays of arguments, where each argument consists of a direction and persuasiveness score (ID 8, ID 9):

$$IA_i = (a_1, a_2, \ldots, a_n) \text{ with } a_j = (d_j, p_j)$$

$$AA_i = (a_{n+1}, a_{n+2}, \ldots, a_{n+m}) \text{ with } a_k = (d_k, p_k)$$

The individual's initial tendency is calculated as the weighted average of their initial argument set (ID 2):

$$T_i^{pre} = \frac{1}{n} \sum_{j \in IA_i} p_j d_j$$

The individual's post-discussion tendency integrates both the initial and additional argument sets (ID 4):

$$T_i^{post} = \frac{1}{n_i^{IA} + n_i^{AA}} \left(\sum_{j \in IA_i} p_j d_j + \sum_{j \in AA_i} p_j d_j \right)$$

The initial aggregate group tendency is the arithmetic mean of all individual initial tendencies (ID 17):

$$\overline{T}^{pre} = \frac{1}{N} \sum_{i=1}^{N} T_i^{pre}$$

The aggregate group tendency post-discussion represents the sum of the initial group tendency and the average impact of the newly exchanged arguments (ID 18):

$$\overline{T}^{post} = \overline{T}^{pre} + \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in AA_i} p_j d_j$$

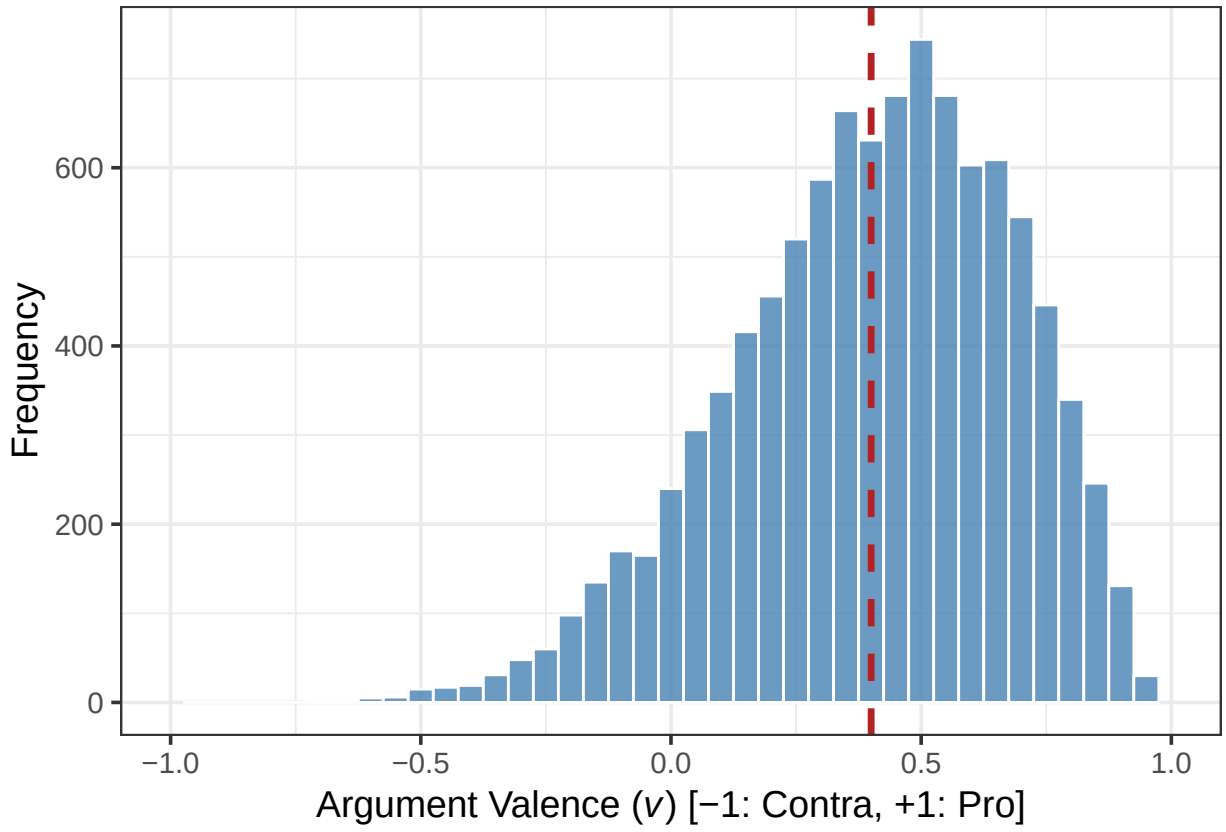
Implementation Deviations. To ensure mathematical stability and behavioral realism within the simulation environment, two structural deviations from the strict theoretical verbalization were implemented.

Deviation 1: Generative Approach to Argument Pools. The formal model abandons the use of static input vectors for argument sets. Instead, it dynamically samples arguments from a finite, simulated pool. The parameters of this pool are strictly

manipulable, specifically its absolute size (total number of available arguments) and a group bias parameter (μ). This bias parameter defines the probability distribution of argument valence, reflecting the tendency of groups to systematically share and encounter information that aligns with a specific directional predisposition. Crucially, both the initial arguments (IA_i) assigned to an agent prior to interaction and the additional arguments (AA_i) encountered during the simulated group discussion are drawn from this identical pool.

Figure 4

Distribution of Argument Valence in the Generated Pool



Note. Histogram showing the scaled Beta-distribution centered around the bias parameter.

As illustrated in Figure 4, the generative pool enforces a structural bias in the distribution of argument valence.

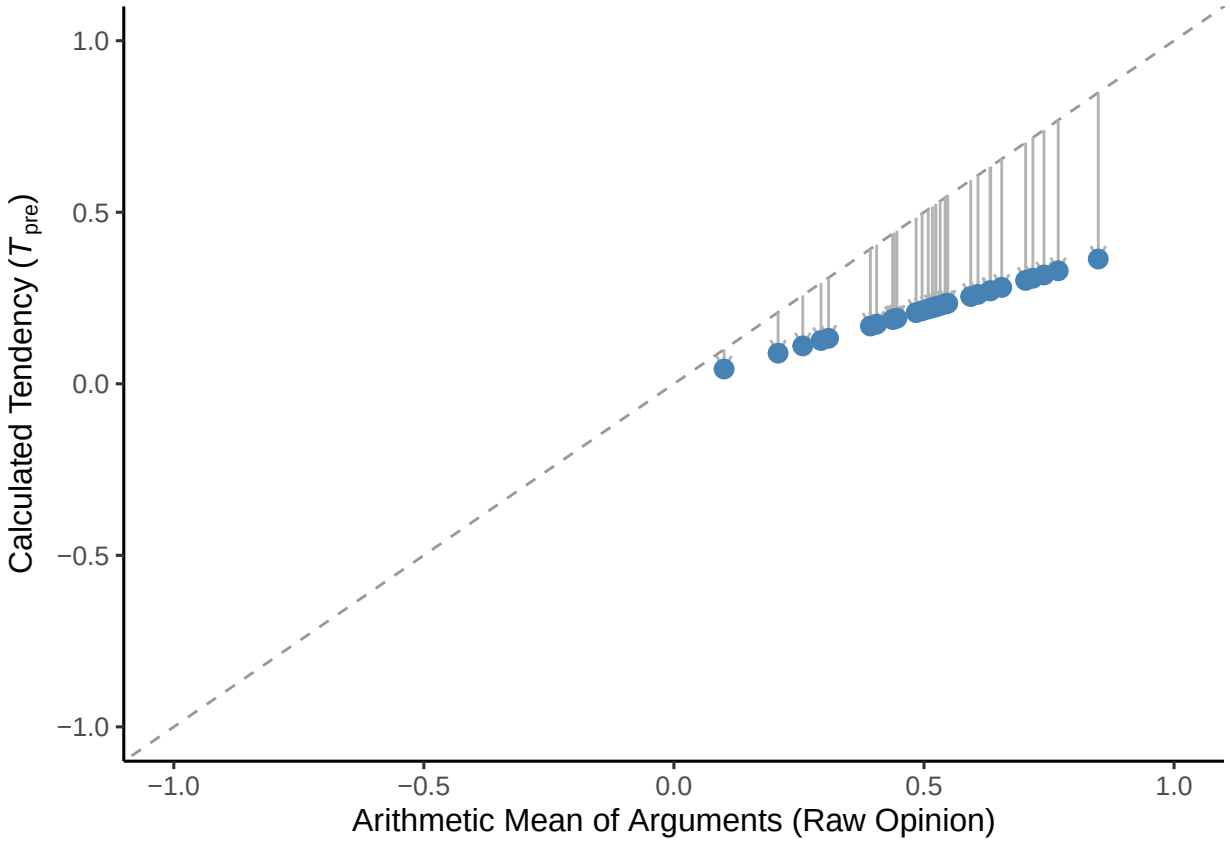
Deviation 2: Calculation Logic and Uncertainty-Dependent Stability. Our baseline mathematical translation of the theory fails to endogenously produce the target phenomenon. Because both initial and additional arguments are drawn from the same biased pool, the statistical integration of AA_i merely reduces the variance of opinions across the group (regression toward the pool mean). It does not naturally induce directional choice shifts or group polarization. To systematically force the phenomenon's emergence, a mathematical dampening factor (C) was introduced into the denominator of the tendency equations:

$$T_i = \frac{1}{n + C} \sum_{j=1}^{n} p_j d_j$$

Psychologically, the C parameter operationalizes uncertainty-dependent stability. When an agent possesses a critically low volume of information (a small number of arguments n), the denominator heavily damps the calculated tendency toward the neutral zero-point. As the individual acquires more arguments during discussion, the relative impact of C diminishes, allowing the opinion to un-damp, stabilize, and reach extreme values. While mathematically necessary to produce the shift within this specific formalization, the psychological realism and theoretical validity of this dampening assumption remain open to debate.

Figure 5

Impact of Uncertainty on Initial Opinion Formation



Note. Comparison of raw arithmetic mean versus damped tendency with uncertainty factor C.

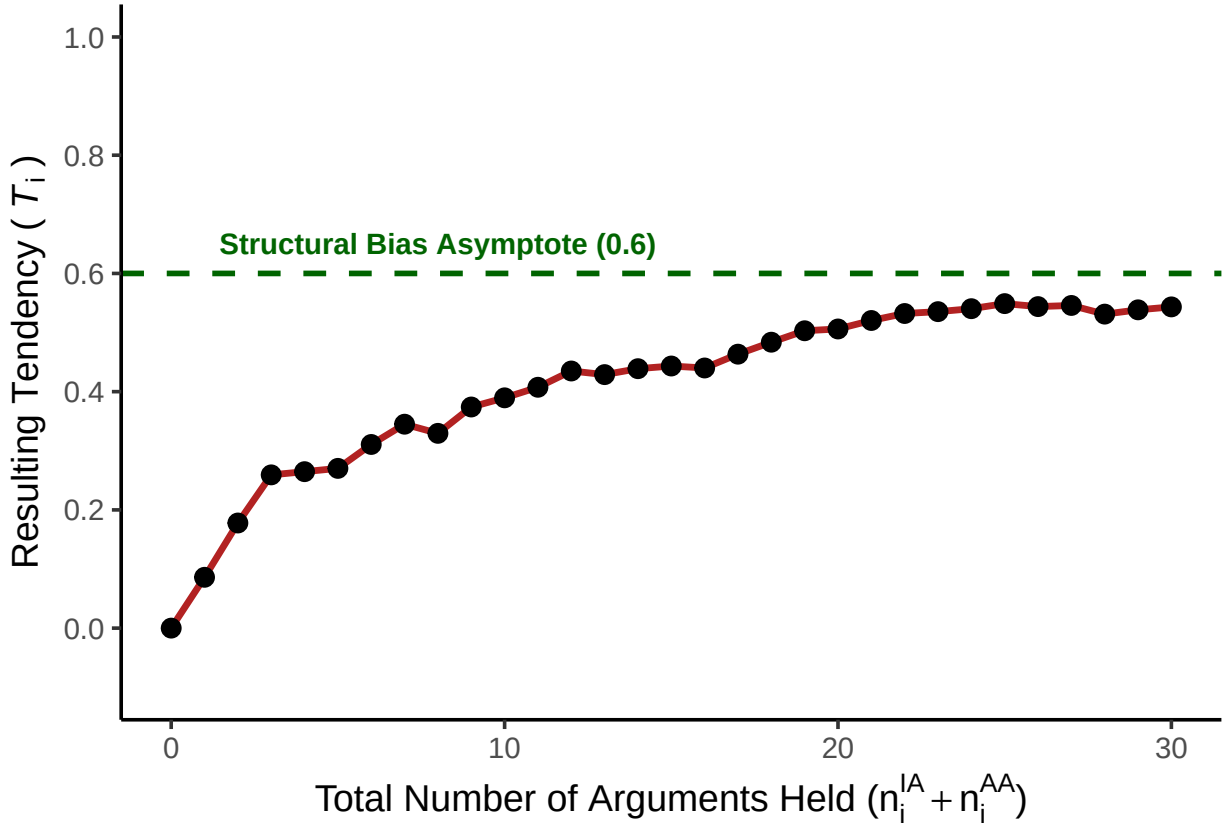
Figure 5 demonstrates how the inclusion of the C parameter pulls initial opinions with low argument counts toward a moderate state, creating the necessary mathematical baseline for subsequent extremization.

System Behavior and Phenomenon Emergence. To evaluate the combined computational architecture at a theoretical level, the behavior of the complete formal model must be analyzed. This overarching mechanism takes the complete set of defined

parameters as inputs, including the pool size (N) pool bias (μ), the number of initial and additional arguments (n_i^{IA}, n_i^{AA}), and the dampening factor (C), and computes the continuous evolution of opinion tendencies.

Figure 6

Theoretical Behavior of the Complete Formal Model



Note. Update trajectory of an individual's tendency as it approaches the pool bias asymptote.

Figure 6 visualizes this systemic transfer function by mapping the trajectory of a single agent. Because both initial arguments (n_i^{IA}) and additional arguments (n_i^{AA}) are

drawn from the identical generative pool, a mathematical distinction between these two sets is unnecessary for this theoretical analysis. The plot demonstrates the core mathematical mechanism across the continuous integration of all arguments. Continuous exposure to a structured pool bias systematically overrides the initial uncertainty dampening. This process drives the individual's tendency (T_i) asymptotically toward the pool bias.

Crucially, this isolates the micro-level mechanism. It confirms that the mathematical architecture successfully generates a directional choice shift for a single agent. It does not display aggregate group polarization. To verify if these isolated micro-level shifts compound into the macro-level target phenomenon of group polarization, an empirical simulation involving multiple agents is required in the subsequent evaluation phase.

Individual Contribution: Dynamic Confirmation Bias Mechanism. An alternative computational approach is proposed to internally model the concept of argument 'validity' without relying on the external dampening factor C. This approach replaces the static arithmetic weighting with a dynamic confirmation bias mechanism. Arguments aligning directionally with an agent's prior tendency receive higher weighting, while dissonant arguments are systematically discounted proportional to the strength of the agent's current opinion.

Simulation data confirms that this mechanism successfully reproduces the group polarization phenomenon natively. The following core R functions define this alternative architecture:

```
create_pool <- function(pool_size, pool_bias) {
  prob_pos <- (pool_bias + 1) / 2
  arguments <- sample(c(-1, 1), size = pool_size, replace = TRUE,
                      prob = c(1 - prob_pos, prob_pos))
  return(arguments)
}
```

```r
initialize_agent <- function(agent_id, pool, n_IA) {
  ia_values <- sample(pool, size = n_IA, replace = FALSE)
  ia_weights <- rep(1, n_IA)
  sum_val <- sum(ia_values * ia_weights)
  sum_w   <- sum(ia_weights)
  t_pre <- if(sum_w == 0) 0 else sum_val / sum_w

  list(id = agent_id, sum_val = sum_val, sum_w = sum_w,
       t_pre = t_pre, t_post = NA_real_, shift = NA_real_)
}

discuss_and_update <- function(agent, pool, n_AA, conf_bias_strength = 1) {
  current_sum_val <- agent$sum_val
  current_sum_w   <- agent$sum_w
  current_t       <- agent$t_pre

  for(k in 1:n_AA) {
    new_arg <- sample(pool, size = 1, replace = TRUE)
    is_consonant <- (new_arg * current_t) >= 0

    if (is_consonant) {
      weight <- 1.0
    } else {
      weight <- 1.0 - (abs(current_t) * conf_bias_strength)
      if(weight < 0) weight <- 0
    }
```

```
    current_sum_val <- current_sum_val + (new_arg * weight)
    current_sum_w   <- current_sum_w + weight
    current_t <- current_sum_val / current_sum_w
  }

  agent$t_post  <- current_t
  agent$shift   <- agent$t_post - agent$t_pre
  agent$sum_val <- current_sum_val
  agent$sum_w   <- current_sum_w

  return(agent)
}
```

A standardized execution of this simulation paradigm (`run_simulation(n_agents = 10, pool_size = 1000, pool_bias = 0.5, n_IA = 6, n_AA = 20, conf_bias_strength = 0)`) demonstrates robust choice shifts. A paired t-test between the generated T_i^{pre} and T_i^{post} arrays, alongside a comparative density plot detailing the temporal shift in attitude extremization, statistically validates the capacity of the dynamic confirmation bias model to reliably output the target phenomenon.

Results

Discussion

- Eigenleistung um Validity noch einfügen (sowohl bei Methoden als auch bei Results, aber jeweils nur kurz)
- Punkt 7 und 8 machen (Results und Discussion) - Kriterien und mit Gemini erstellte Grobstruktur checken

- Tables in den Appendix packen
- Verwendete Materialien angeben: R, RStudio, Gemini version xx (zum Formulieren und zur Unterstützung beim Coden)
- Nochmal jpg in der Gruppe checken mit ToDos für die Arbeit + Nochmal Kriterien auf Felixs homepage checken + nochmal checken, dass der ganze Code direkt/indirekt in die Arbeit gefunden hat
- Kürzen!!
- Abstract schreiben

References

Burnstein, E., & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology*, *13*(4), 315–332. https://doi.org/10.1016/0022-1031(77)90002-6

Cronbach, L. J., & Shapiro, K. (1982). *Designing Evaluations of Educational and Social Programs*. Jossey-Bass.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151. https://doi.org/10.1037/0022-3514.50.6.1141

Leising, D., Grenke, O., & Cramer, M. (2023). Visual Argument Structure Tool (VAST) Version 1.0. *Meta-Psychology*, *7*. https://doi.org/10.15626/MP.2021.2911

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135. https://doi.org/10.1037/h0027568

Pedraza, L., Saintier, N., Pinasco, J. P., Balenzuela, P., & Anteneodo, C. (2025). Analytical insights from a model of opinion formation based on persuasive argument theory. *Physical Review E*, *112*(1), 014312. https://doi.org/10.1103/j5z4-5ry9

Van Lissa, C. J., Peikert, A., Ernst, M. S., Dongen, N. N. N. van, Schönbrodt, F. D., & Brandmaier, A. M. (2026). To Be FAIR: Theory Specification Needs an Update. *Perspectives on Psychological Science*, 17456916251401850. https://doi.org/10.1177/17456916251401850