PAULINA ŻELEWSKA, BARBARA KONAT

THE LANGUAGE OF AFFECT HEURISTIC IN ONLINE DISCOURSE: A CORPUS STUDY WITH HUMAN AND LLM CODERS

ABSTRACT. Paulina Żelewska, Barbara Konat, *The Language of Affect Heuristic in Online Discourse: A Corpus Study with Human and LLM Coders*, edited by Barbara Konat, Cristián Santibáñez, „Człowiek i Społeczeństwo" vol. LXI: *Emotion in Argumentation*, Poznań 2026, pp. 103–118, Adam Mickiewicz University. e-ISSN: 2956-5243, ISSN 0239-3271, https://doi.org/10.14746/cis.2026.61.8.

This paper investigates the affect heuristic in online argumentative discourse through a corpus-based study combining manual and automatic annotation. Drawing on research in argumentation theory, psychology, and decision science, the study approaches affect not as an external addition to reasoning but as a recurring component of evaluative judgment. The analysis focuses on discussions of climate change and artificial intelligence collected from Reddit and X (formerly Twitter), domains characterised by uncertainty, risk perception, and public controversy.

The study employs a bottom-up annotation methodology in which four human annotators identify instances of affect heuristic and related cognitive biases in a corpus of more than 30,000 posts and comments. Inter-annotator agreement is assessed using Fleiss' κ, Cohen's κ, Gwet's AC1, and weighted F1 measures. In a second stage, the same annotation scheme is applied to GPT-4o, treated as a constrained fifth annotator operating within predefined categories and probabilistic classification rules.

The results show that the affect heuristic is the most frequent heuristic pattern in the corpus, occurring more often than confirmation bias, availability heuristic, or representativeness heuristic. Although traditional κ coefficients remain low because of category imbalance, agreement measures robust to prevalence effects indicate substantial consistency among annotators. The automatic annotation stage reveals partial alignment between human and model judgments, while also exposing systematic discrepancies in the model's distribution of categories.

A lexical and discursive analysis further demonstrates that affect heuristics do not necessarily manifest through explicit emotion vocabulary. Instead, they frequently appear through evaluative framing, practical reasoning under uncertainty, and subtle stance-taking related to collective action and future-oriented judgment. The findings contribute to empirical research on emotional processes in argumentation and demonstrate how affective reasoning can be operationalised and studied through combined qualitative and computational methods.

Keywords: affect heuristic, emotions in argumentation, online discourse, argumentation in social media, cognitive biases, large language models (LLMs)

Paulina Żelewska, Adam Mickiewicz University in Poznań, Faculty of Psychology and Cognitive Sciences, Szamarzewskiego 89AB, 60-568 Poznań, WSB Merito Gdańsk University, Faculty of Social Science and Humanities, Aleja Grunwaldzka 238A, 80-266 Gdańsk, e-mail: zelewska.p@gmail.com, https://orcid.org/0009-0000-6740-135X

Barbara Konat, Adam Mickiewicz University in Poznań, Faculty of Psychology and Cognitive Sciences, Szamarzewskiego 89AB, 60-568 Poznań, e-mail: bkonat@amu.edu.pl, https://orcid.org/0000-0003-2370-4636

Introduction

The traditional opposition between "rational" and "emotional" judgment has become increasingly difficult to defend. In argumentation theory, work since Douglas Walton has shown that emotional appeals in certain context can be legitimate components of persuasive reasoning (Walton, 1992). In psychology and decision science, converging evidence indicates that affective responses often guide good judgment, especially under uncertainty or limited deliberation (Damasio, 1994; Kahneman, 2011; Slovic et al., 2007; Zajonc, 1980). Taken together, these lines of research support a shared conclusion: evaluation and reasoning are routinely intertwined with affect.

This paper starts with that interdisciplinary premise and examines one specific discursive expression of this mechanism: the affect heuristic as it manifests in online discourse. The affect heuristic is commonly understood as a process in which people rely on rapid affective evaluations as informational cues when forming judgments, including judgments about risks and benefits (Kahneman, 2011; Slovic et al., 2007). Individuals often treat positive or negative affect as a proxy for value and probability, which can systematically shape perceived danger, advantage, and responsibility (Finucane et al., 2000). Importantly for the present study, affect-driven judgment does not require overtly emotional language; it can operate through evaluative cues and stance-taking that guide interpretation without explicit emotion terms.

Online platforms provide a particularly revealing context for studying affect-based judgment. Argumentative exchanges on social media are typically fast, attention-limited, and socially embedded, which makes reliance on heuristic evaluation plausible. At the same time, public discussions of climate change and artificial intelligence are domains where risk perception is central and where affective appraisals are likely to influence argumentative positions

and the interpretation of evidence. This creates a natural setting in which affect heuristics should be observable and quantifiable. Our question in this study can be therefore formulated as follows: How does the affect heuristic manifest in social media discourse on risk-laden topics?

The present study contributes an empirical, corpus-based account of affect heuristics in online discourse by combining (i) a bottom-up, iterative manual annotation procedure and inter-annotator agreement analysis, and (ii) an automatic annotation stage using a large language model. The goals are (a) to assess whether affect heuristics can be reliably identified in naturally occurring posts and comments, despite the interpretive complexity of affective phenomena; (b) to compare their distribution to other heuristics and cognitive biases in the same dataset; and (c) to examine how closely LLM-based outputs align with human judgments when the task is framed as probabilistic classification.

Affect in the present study includes evaluative orientations manifested indirectly through discourse, including stance-taking and practical reasoning under uncertainty. Human annotators and the language model therefore identify arguments that appear to be guided by affect heuristics even when they do not contain explicit emotion vocabulary.

Conceptually, this approach is consistent with an interactional and empirically grounded view of emotional effects in argumentation (Authors): affect is treated here not as an optional "irrational" layer but as a recurring component of evaluative reasoning that becomes visible in discourse. Methodologically, the paper demonstrates how such a phenomenon can be operationalized through annotation guidelines and evaluated both with agreement statistics and with comparative human–model analysis, following our own recent studies in methodology of automated analysis (Author).

Materials & methods

The methodological choices adopted in this study follow directly from the theoretical assumptions outlined in the Introduction. If affect is understood as a recurrent component of evaluative reasoning rather than as an external or "irrational" addition to argumentation, then it must be examined where such reasoning naturally occurs: in situated, everyday discourse. From this perspective, social media platforms offer access to large volumes of spontaneous argumentative interaction in which affective evaluations are articulated, negotiated, and normalised in public view.

At the same time, studying affect in natural language requires methodological tools that can bridge qualitative interpretive analysis and quantitative validation. The present approach therefore combines manual pragmatic annotation with formal agreement measures and a comparative human–model analysis. This design allows affect heuristics to be operationalised through explicit annotation guidelines while retaining sensitivity to context, interpretation, and discourse-level phenomena. The data consists of discourse sample collected from the social media platforms Reddit and X (formerly Twitter). Posts and comments are selected to capture discussions on climate change and artificial intelligence, with a balanced representation of both topics and both platforms.

The full corpus contains more than 30,000 instances. Each annotator works with 7,281 units from Reddit and 2,717 from X, with both topics represented equally. Data from X is gathered in April 2024 using the keywords *artificial intelligence*, *climate change*, and *climate crisis*. The Reddit portion, collected in March 2024, comes from the subforums r/singularity, r/aiwars, r/climateskeptics, and r/climatechange.

To evaluate inter-annotator agreement, a portion of the data is annotated by all members of the team of four annotators. This subset, representing approximately 30% of the entire corpus, is embedded into each annotator's individual dataset so that it does not stand out during the annotation process. This agreement subset is used to measure annotation consistency and later serves as the basis for the automatic annotation stage.

The annotation procedure follows a bottom-up methodology developed by Marcin Koszowy et al. (2022) for analysing complex phenomena expressed in natural language. The work was carried out across several iterative rounds, with the scheme created in the first round serving as the basis for subsequent stages. This iterative structure enabled the continuous identification of error sources and refinement of ambiguous definitions in the scheme, ultimately improving the agreement of the annotations.

Following the procedure used for human annotators, the same annotation scheme and instructions for identifying instances of the affect heuristic are applied to the GPT-4o model on the same 30% subset of the corpus; its output is treated as that of a fifth annotator. Comparing the model's output with human annotations allows us to assess their alignment and identify the model's weaknesses, particularly in recognising less frequent cognitive biases. Agreement measures are then calculated, and discrepancies are examined using confusion and contingency matrices.

In the automatic annotation stage, the language model is not treated as an autonomous interpreter or generator of analytic categories, but as a constrained

annotator operating under conditions analogous to those imposed on human coders. Rather than asking the model to "interpret" texts freely or to produce explanations, the task is explicitly limited to assigning likelihood estimates to a closed and predefined set of heuristic and bias categories.

This restriction is methodologically crucial. Large language models have a well-documented tendency to elaborate, paraphrase, or introduce new conceptual distinctions when allowed to respond freely. To avoid this, the prompt is formulated as a strict instruction that mirrors the human annotation protocol: the model receives a fixed inventory of labels and is explicitly prevented from introducing new categories, redefining existing ones, or producing discursive commentary. Its output is limited to numerical probability values assigned to each category. In this sense, the model does not "learn" new notions of bias during annotation but applies predefined theoretical constructs to the text, just as human annotators do.

Using a language model in this way should therefore not be understood as delegating interpretation to an artificial agent, but as testing whether a computational system can approximate human judgments when both are constrained by the same conceptual framework and annotation rules. The probabilistic output format further reflects the graded and often ambiguous nature of heuristic identification, allowing the model to express uncertainty in cases where even human annotators diverge.

The model's main task is to assign a probability score to each category, reflecting how likely it is that a given heuristic or cognitive bias is present in the input. Two additional categories are included: *no bias*, used when the model does not detect any cognitive bias, and *unclear*, used when the content cannot be classified with certainty. By generating probability estimates instead of binary 0–1 labels, the model is able to offer a more fine-grained interpretation of the data and better capture borderline cases that may be ambiguous even for human annotators.

Results

Results for human annotators

Across the full dataset, the affect heuristic is by far the most frequently annotated category, with 2,662 instances. Other heuristics and cognitive biases occur substantially less often, including confirmation bias (1,030) and the availability heuristic (616), while representativeness heuristic (265) and all-or-nothing

thinking (244) are comparatively rare. This distribution identifies the affect heuristic as the dominant form of heuristic reasoning in the corpus.

A closer look into the 30% reference set of the full corpus, annotated independently by all four coders, reveals the differences between human annotators. Table 1 shows the number and percentage of found cases of each heuristic and bias type in this subset. It is clear, that affect heuristics constitutes the most numerous category, however there are notable differences between annotators. Ann2 provides the highest number of cases, with affect heuristic making up 66.47% of their total. Ann3 and Ann4 show a more even distribution. For Ann3, confirmation bias (25.62%), availability heuristic (22.21%), and representativeness heuristic (19.19%) appear in similar proportions. Ann4 also relies most often on affect heuristic (39.09%), while the remaining categories range between 11% and 24%. Ann1 contributes the fewest cases overall, with confirmation bias as the most frequently assigned type (30.28%), followed by all-or-nothing thinking (25.69%) and affect heuristic (22.94%).

Table 1. Distribution of annotated heuristics and cognitive biases by individual coders in the reference set with percent of each coder's total marked cases

Coder	Bias	Count	Percent of coder's total
Ann1	Affect heuristic	25	22.94
	Availability heuristic	15	13.76
	Representativeness heuristic	8	7.34
	Confirmation bias	33	30.28
	All-or-nothing thinking	28	25.69
Ann2	Affect heuristic	594	62.86
	Availability heuristic	121	12.80
	Representativeness heuristic	15	1.59
	Confirmation bias	199	21.06
	All-or-nothing thinking	16	1.69
Ann3	Affect heuristic	167	31.04
	Availability heuristic	107	19.89
	Representativeness heuristic	66	12.27
	Confirmation bias	119	22.12
	All-or-nothing thinking	79	14.68
Ann4	Affect heuristic	42	45.65
	Availability heuristic	10	10.87
	Representativeness heuristic	3	3.26
	Confirmation bias	23	25.00
	All-or-nothing thinking	14	15.22

Source: own research.

Taken together, the results demonstrate both that the annotation of affect heuristics is empirically feasible and that the affect heuristic plays a central role among the identified heuristic patterns in online discourse.

Table 2. Values of Fleiss' κ and Gwet's AC1 for overall bias presence and individual heuristics and cognitive biases

Type of heuristic or bias	Fleiss' κ	Gwet's AC1
Bias presence	0.0309	0.9442
Affect heuristic	0.0556	0.8196
Availability heuristic	0.0861	0.9498
Representativeness heuristic	0.0288	0.9810
Confirmation bias	0.0745	0.9262
All-or-nothing thinking	0.0721	0.9737

Source: own research.

The agreement subset shows the same pattern as the full dataset. The affect heuristic remains the most frequently assigned category, indicating that its prominence is not an artifact of partial annotation but persists under full independent double coding.

Inter-annotator agreement is assessed using Fleiss' κ, Cohen's κ, Gwet's AC1, and the weighted F1 score. For the affect heuristic, Fleiss' κ is low (0.0556), while Gwet's AC1 indicates substantial agreement (0.8196). This discrepancy reflects the well-known sensitivity of κ to uneven category distributions, particularly in cases where one category is dominant.

Across annotator pairs, the same pattern is observed: Cohen's κ remains low, whereas AC1 and weighted F1 scores are consistently higher, as visible in Table 2. These results show that, despite the subjectivity typically associated with affective phenomena, the affect heuristic can be identified with meaningful and stable agreement when agreement measures robust to prevalence effects are applied.

Results for automatic annotation

Automatic annotation is performed on the agreement subset, which represents 30% of the data annotated by all members of the team. This subset contains posts from both topics and both platforms that have already been coded manually and previously served to assess inter-annotator agreement. In the present stage of the study, it functions as a reference point for determining whether the model reaches a level of agreement comparable to that

observed among human annotators, and whether it identifies the same types of cognitive biases as the annotators.

Unlike the human annotators, the model does not assign a single label to each post. Instead, it returns a probability score for every category. For each entry, the model estimates the likelihood that the text reflects one of the following categories: affect heuristic, availability heuristic, representativeness heuristic, confirmation bias, all-or-nothing thinking, or no cognitive bias. In some cases, the model assigns similar probability values to several categories, resulting in ambiguous outputs. To capture this ambiguity, an additional *Unclear* category is introduced. This label is applied when two conditions are met simultaneously: (i) the standard deviation of the probability scores across categories is below 0.093, and (ii) the probability range does not exceed 0.20. These thresholds are determined based on posts whose probability spread falls below the 25th percentile (0.35); for this subset, the mean standard deviation is 0.093 and the mean range is 0.24. In total, 125 posts (4.17%) are classified as ambiguous.

The overall distribution of labels assigned by the model is presented in Table 3. The most frequently identified category is all-or-nothing thinking (36.05%), followed by no cognitive bias (26.24%) and confirmation bias (14.84%). The remaining categories appear less often: affect heuristic at 11.30%, availability heuristic at 6.20%, and representativeness heuristic at 5.70%.

Table 3. Distribution of heuristics and cognitive biases assigned by the model in the agreement subset

Type of heuristic or bias	Count	(%)
Affect heuristic	339	11.30
Availability heuristic	186	6.20
Representativeness heuristic	171	5.70
Confirmation bias	445	14.84
All-or-nothing thinking	1081	36.05
No cognitive bias	787	26.24
Unclear	125	4.17

Source: own research.

Agreement of the automatic annotation is assessed in two ways. In the first approach, the model is treated as a fifth annotator alongside the four human annotators (Ann1, Ann2, Ann3, Ann4). Fleiss's κ and Gwet's AC1 are calculated to measure agreement when all annotators are considered together. In the second approach, the model is compared separately with

each human annotator. This makes it possible to determine which annotator the model is most closely aligned with and whether its level of agreement is similar across individuals.

Table 4 presents the values of Fleiss' κ and Gwet's AC1 for overall bias presence as well as for each individual heuristic or cognitive bias. The highest AC1 values are obtained for bias presence (0.9442), the representativeness heuristic (0.9650), and the availability heuristic (0.9446). For the affect heuristic, the model reaches a moderate level of agreement, with an AC1 score of 0.8447 and a low Fleiss' κ of 0.0736. The contrast between the high AC1 and low κ indicates that the model is sensitive to the uneven distribution of categories. Because the affect heuristic appears relatively infrequently in the dataset, the imbalance lowers the κ coefficient even in cases where the actual agreement is high.

Table 4. Values of Fleiss' κ and Gwet's AC1 for overall bias presence and individual heuristics in the agreement subset

Type of heuristic or bias	Fleiss' κ	Gwet's AC1
Bias presence	–0.0106	0.9442
Affect heuristic	0.0736	0.8447
Availability heuristic	0.0813	0.9446
Representativeness heuristic	0.0189	0.9650
Confirmation bias	0.0610	0.8919
All-or-nothing thinking	–0.0361	0.8182

Source: own research.

Language of affect heuristics

Lexis of affect heuristics

To explore the lexis associated with the affect heuristic, we employ word clouds as an exploratory descriptive tool. A word cloud is a visual representation of lexical frequency in which words occurring more often in a given set of texts are displayed more prominently. In the present study, word clouds are generated from the subset of posts annotated by human annotators as instantiating the affect heuristic, after standard preprocessing steps such as lowercasing and the removal of high-frequency function words that do not carry interpretive weight. The purpose of this visualization is not to provide a statistical model of affect, but to offer an intuitive overview of the

vocabulary that characterises affect-based evaluative reasoning in discourse. For the analysis of the affect heuristic, word clouds are particularly useful because they allow us to examine whether affect manifests through overtly emotional language or through more subtle lexical choices that reflect evaluation, stance, and orientation toward shared concerns.

The definition of the affect heuristic assumes that people often rely more on affective impressions than on objective analysis when forming judgments, as outlined in Introduction. One might therefore expect this heuristic to appear through emotionally charged language or through statements expressed in an extreme tone, however word clouds from our corpus allow for more complex context exploration. First, Figure 1 a. and b. show the most frequent terms in affect heuristics across all topics and platforms (the first one is the raw cloud, the second – with most frequent words, i.e. stopwords, removed). As we can see, those clouds do not contain typical items that could be expected based on handbook examples, with emotion names (fear, joy) or emotional expressions (afraid, happy). What we can see instead, are clouds of abstract concepts (such as "people," "human," "world") in combination with verbs related to cognitive states ("think," "want," "need"). These are typical for practical reasoning scheme (Walton, 1992) in argumentation, that is the type of reasoning where conclusion is a call to action. From the lexical choices by the speakers in our discourse we can see that they are thinking about action, and a large scale one, about what we, as "people," as "the world" should or want to do. Moreover, practical reasoning, especially in uncertain conditions can be also a presumptive reasoning (i.e. based on incomplete premises and missing information) Walton indicated, is the one where emotional appeals are frequently used, and can be considered adequate and not fallacious. Our annotators noted that when marking the presence of affect heuristics – the speakers in the discourse have obviously incomplete information the future in AI or climate change is uncertain, therefore their call to actions are based on approximate and evaluative judgements, making the affect heuristics intertwined with practical reasoning.

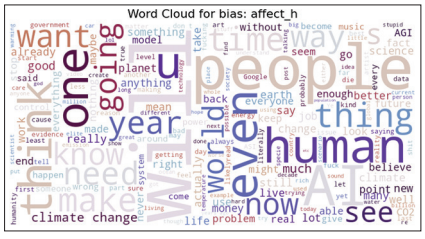
When comparing the affect heuristic across topics, clear differences emerge between climate-related discourse and discussions on artificial intelligence. In posts about climate change, scientific rather than alarmist framing dominates, and the language focuses more on understanding the issue than on appealing to emotion. Words such as *problem*, *issue*, *wrong* and *believe* point to ongoing debate and climate scepticism rather than narratives rooted in fear. In contrast, AI-related discussions include terms like *model*, *artist*, *Google*, *technology* and *future*, which suggest that conversations centre

on the social impact of AI, particularly in relation to work, creativity and technological development. Importantly, neither topic contains vocabulary that signals strong fear or anger, which challenges the assumption that the affect heuristic always appears through highly emotional language.

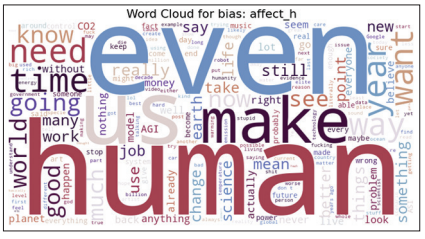
A particularly interesting term is *human*, which appears frequently in AI discussions on *Reddit*. Its presence may signal an emotional dimension in the discourse by emphasising human qualities such as judgment, creativity and morality. When users invoke *human* in AI debates, they often highlight the distinction between people and machines, which can evoke affective reactions such as scepticism toward AI in areas traditionally associated with humans.

An analysis of the affect heuristic at the platform level shows distinct differences between *Reddit* and *X* (formerly Twitter), both in emotional tone and in language use. On *Reddit*, the word *even* stands out (Figure 1e), functioning as an intensifier that strengthens the emotional impact of a statement by expressing surprise, frustration or disbelief. The language of posts and comments on *X* is considerably more confrontational and often openly hostile (Figure 1f). Words such as *liar*, *Jew*, *evil* and *racist* indicate that the discussions frequently intersect with ideological and political conflicts. The presence of commonly recognised vulgar or offensive terms, including *shit* and *fuck*, shapes a direct and highly emotional tone of communication. This type of expression aligns with the nature of the affect heuristic, although it raises the question of whether it reflects the heuristic itself or the inherent dynamics of the platform.

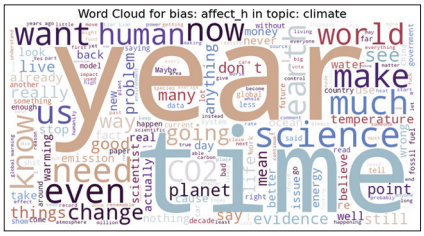
Overall, the analysis suggests that the affect heuristic does not always manifest through overtly emotional language. It may appear through subtle emotional cues, which are particularly visible in posts from *Reddit*. The way in which this heuristic emerges in online discourse seems to depend on the communication style typical of each platform, ranging from the more reflective tone on *Reddit* to the emotionally intense reactions that are common on *X*.



(a) Word cloud for the affect heuristic in the full dataset

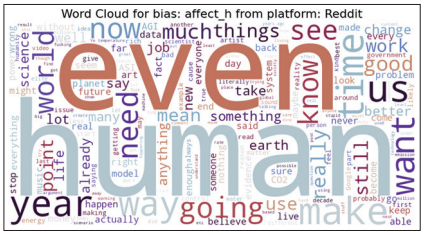


(b) Word cloud for the affect heuristic in the full dataset after removing additional stopwords



(c) Word cloud for the affect heuristic in intelligence discussions

(d) Word cloud for the affect heuristic in climate change discussions artificial



(e) Word cloud for the affect heuristic in Reddit posts

(f) Word cloud for the affect heuristic in X (Twitter) posts

Figure 1. Comparative visualisations of word clouds for the affect heuristic across different contexts. The six subfigures show: (a) the full dataset, (b) the version after removing additional stopwords, (c) climate change discussions, (d) artificial intelligence discussions, (e) Reddit posts, and (f) posts from X (Twitter).

Source: own research.

Discursive manifestation of affect heuristics

After presenting the characteristic lexemes of the affect heuristic, this section illustrates how these cues were annotated in practice. Although the annotation scheme is grounded in theoretical assumptions and explicit definitions, annotation remains an interpretive activity, shaped by individual judgments and sensitivities. The examples discussed below come from the agreement subset (i.e. the narrow set of cases with perfect agreement) of the corpus (3,000 posts) and are used to show both cases of full annotator consensus and cases where judgments diverged.

We begin with instances of *perfect agreement*, that is, posts for which all four annotators independently identified the affect heuristic. Such cases are rare in the dataset, but when they occur, they typically involve a clearly

negative evaluative stance and explicit expressions of concern or opposition. The following examples come from both Reddit and X and cover both thematic domains analysed in the study.

(1) a. I worry about a big heatwave that kills a hundreds of millions/half a billion in a few days, as it rolls across Pakistan, India, Bangladesh, and maybe Indonesia. When it hits that wet bulb mark, and people can't cool down. It won't take a year, it'll take 2 or 3 days.
b. Artificial intelligence! It is a deadly technology that will put ordinary hardworking people in great danger in the future! This technology will make capitalists more profitable and ordinary workers will lose their jobs. #AI #disaster #Unemployment
c. I honestly dont want ai to replace the creativeness of humans, but it will. sad. https://t.co/Z0iq0uYCzP

Cases presented in Example (1) illustrate typical emotional expressions of fear (a, b) and sadness (c). We observe clear lexical indicators, such as "kill" or "danger," along with an explicit emotion term ("sad"). On the surface, these cases function primarily as expressions of emotional release, with no explicit call to action. However, each example introduces a future-oriented scenario: the speakers reason about what may happen, and this reasoning is motivated by affect heuristics, thus instantiating the case of practical reasoning.

Let us now analyze cases of *partial agreement*, in which three out of four annotators identified the same heuristic or combination of heuristics.

(2) a. @edgarrmcgregor Edgar, here are some words to describe the current climate crisis: „There is no climate crisis. „Don't be duped by devious, freedom-hating and control-hungry slimeballs.

Example (2) presents a very interesting case of emotional appeal in argumentation combined with affect heuristics (as marked by 3 out of 4 annotators). First, the speaker is making an argument, with the premise "There is no climate crisis," leading to conclusion "You should not be duped," proceeding with offensive words with strong metaphorical element. So it is practical reasoning in terms of the structure (as the conclusion is a call to action, in this case an action a hearer should take). It is also an emotional argument, filled with emotion-eliciting words such as "duped" (why would anyone want to be duped?), "freedom-hating" (how can you listen to someone who hates freedom), not to mention offensive "slimeballs." So just like in cases analysed in previous works on the topic (Author), the speaker is using emotion-eliciting words with the aim of increasing the perceived strength of their argumentation.

But from the perspective of affect heuristics, something else, interesting is happening in Example (2). As our annotators instructed for observing affect heuristics noticed, the speakers judgement here is also influenced by their own emotion. Because they evaluate people showing evidence for climate crisis so negatively, we can assume this influences the speakers own judgements, inability to look at evidence and to accept or reject things more rationally. This example is a complex case showing how discourse manifestations can be a window into both speakers rhetorical strategies, but also showing how those reveal unconscious bias on their own.

Let us now move to other typical cases, showing similar patterns:

(3) a. That's a tough spot to be in with a good friend. I'm also in southern Ontario and working on doing the opposite: moving a little further north to do some regenerative agriculture and rewilding. Your friend will regret her decision. There aren't many places better suited to our terrible future than southern Ontario. But even here, most people aren't willing to have a frank conversation on the topic.

In this Reddit post on climate change (3), annotators diverged between identifying the affect heuristic, the availability heuristic, both, or no bias. The post combines personal experience, future-oriented concern, and generalisation, making it a borderline case where affective evaluation and experiential salience overlap. We can observe multiple emotion-appealing and emotion-expressing words (marked in bold). The post also shows an emotional see-saw, mixing up positive words (such as "friend") with fear-eliciting ("terrible").

(4) I think it is the only way forward. We will never contain AGI. We must treat it the same way we treat other humans or eventually it will rebel.

Example (4) (Reddit, AI topic) is interesting as a borderline case: it contains explicit expressions of fear alongside categorical claims (e.g. "never contain AGI"), illustrating how affective evaluation can coexist with extreme, dichotomous reasoning. In terms of annotation, it elicited partial agreement between the affect heuristic and all-or-nothing thinking. This shows how original heuristics categorization does not always contain mixed-case items taken from natural language.

Across these cases, the affect heuristic was the category most likely to achieve partial agreement, followed by confirmation bias and the availability heuristic. To compare, representativeness heuristic did not reach the 75% agreement threshold in any example, suggesting that it may be more difficult to identify consistently in short, informal posts.

Conclusion

This paper investigated the affect heuristic in naturally occurring online argumentative discourse, combining manual annotation with automatic annotation using a large language model. The results support three main conclusions.

First, the affect heuristic is identifiable in natural language data. Using a bottom-up, iterative annotation procedure applied to social media discourse, the study shows that affect-based evaluative reasoning can be operationalised and systematically annotated in real-world texts. Despite the conceptual complexity of affective phenomena, annotators were able to apply the category in a consistent manner when supported by explicit guidelines.

Second, and more importantly, the affect heuristic constitutes a substantial proportion of the heuristic and bias patterns observed in the corpus. Across platforms and topics, it was the most frequently annotated category, occurring more often than confirmation bias, availability heuristic, or representativeness heuristic. This finding suggests that affect-driven evaluation is not a marginal phenomenon in online argumentation but a dominant mode of judgment in discussions centred on uncertainty and risk. In terms of the linguistics manifestations of affect heuristics, it operates through practical reasoning oriented toward collective action under uncertainty. Because speakers reason with incomplete information about future states, their calls to action rely on approximate and evaluative judgments rather than on fully specified premises.

Third, the annotation of the affect heuristic can be achieved with reasonable agreement, provided that agreement measures robust to category prevalence are applied. While traditional κ coefficients yielded low values due to distributional imbalance, Gwet's AC1 and weighted F1 scores indicated substantial agreement. At the same time, the results also show that further methodological refinement is needed, both in sharpening operational definitions and in better understanding sources of annotator divergence in borderline cases.

In terms of limitations, using a language model in this study functions as a methodological probe into whether a computational system can approximate human judgments when both are equally constrained by the same theoretical definitions, annotation guidelines, and limits on permissible categories. It should not therefore be understood as delegating interpretation or evaluative authority to an artificial agent. The comparison with automatic

annotation revealed systematic discrepancies between human and model outputs. In particular, the language model identified the affect heuristic less frequently than human annotators and substantially overestimated the presence of all-or-nothing thinking. At this stage, no strong explanatory claims can be made about this pattern. Research on the use of large language models as annotators or classifiers of complex, pragmatically grounded phenomena is still limited, and it remains unclear whether the observed differences reflect properties of the model, properties of the prompt design, or deeper challenges in modelling affective evaluation. The present results should therefore be treated as exploratory and motivate further studies rather than definitive conclusions about model performance.

Finally, the findings should be considered in light of their broader social relevance. The two domains examined in this study – climate change and artificial intelligence – are associated with high levels of perceived risk and uncertainty, and online discussions in these areas both reflect and shape how communities understand and respond to these issues. As shown in the work of Amos Tversky and Daniel Kahneman, everyday judgments and decisions are often guided by fast, heuristic-based processes rather than deliberate analysis. Understanding how affect heuristics operate in public discourse is therefore crucial, not only for theoretical models of argumentation and cognition, but also for grasping how collective attitudes and actions emerge in response to socially consequential risks.

Literature

Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. New York: Avon Books.

Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Macmillan.

Koszowy, M., Budzynska, K., Pereira-Fariña, M., Duthie, R. (2022). From theory of rhetoric to the practice of language use: The case of appeals to ethos elements. *Argumentation*, *36*(1), 123–149. http://doi.org/10.1007/s10503-021-09564-0

Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G. (2007). The affect heuristic. *European Journal of Operational Research*, *177*(3), 1333–1352.

Walton, D. (1992). *The Place of Emotion in Argument*. University Park, PA: The Pennsylvania State University Press.

Zajonc, R.B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175. http://doi.org/10.1037/0003-066X.35.2.151