IDENTIFYING AND MITIGATING INHERENT BIAS IN GPT 4.0 USING DIALECTICAL METHOD

Alanas Petrauskas*, alanas196560@gmail.com, Evaldas Taroza, evaldas@taroza.lt, Dialexity Inc., Marcinkeviciaus g. 17-15, Vilnius 08433, Lithuania

Abstract

A novel dialectical method identifies and mitigates GPT's inherent biases, which typically manifest as exaggerated outcomes and a 'quick fix' mentality. This method introduces the concept of ethical-semantic relations by equating the positive and negative aspects of theses and antitheses with inherent goals, risks, and obligations. According to GPT's own assessments, the outcomes derived from this method are more accurate than without its application. GPT's occasional failures to achieve dialectic synthesis and accurate interpretation of some simple statements are also explored.

Keywords: Dialectical AI, XAI, Moral AI, Dialexity

Introduction

Dialectics offers significant potential for validating AI methods due to its universality and influence over all faculties of human intelligence (see e.g. [1-3]). This advantage was historically difficult to harness due to its abstract nature, but this can now be changed due to the semantic-ethical relations exemplified in Table 1. Here, T represents the major thesis of a given text (what we really say), A represents its major antithesis (what would our opponent say), while (+) and (-) denote their positive and negative aspects respectively. These interrelations provide a universal framework for generating moral maxims that are useful in conflict resolution and negotiations [4], fostering the development of both Explainable and Ethical AI (as defined in [5] and [6] respectively).

Table 1. Relations between dialectical elements

Type	Statement	T	T+	T-	A	A+	A-
1.1	Positive (constructive) side of		T	-		A	-
1.2	Negative (destructive) side of		-	T		-	A
2.1	Overdevelopment of		-	T		-	A
2.2	Underdevelopment of		-	A+		-	T+
3.1	Complimentary to		A+	A-*		T+	T-*
3.2	Contrary/Opposite to (A(X))	A	A-	A+	T	T-	T+
4.1	Inherent Goal of	T-	T	-	A-	A	-
4.2	Inherent Risk of			T			A
4.3	Implied Obligation of	-	A	-		T	
4.4	Subsequent Risk of			A			T

* Either complimentary to or following after

The first type of relations (1.1 and 1.2) equates the positive aspects of thesis and antithesis to constructivity, and the negative aspects to destructivity. The second type (2.1 and 2.2) links the negative sides to overdeveloped (exaggerated) and underdeveloped (suppressed) forms of A and T. The third type (3.1-3.2) defines the complementarity (mutual enhancement) between like-signed T and A (rule 3.1), and the incompatibility (mutual suppression) between differently signed T and A (rule 3.2). These relationships can be expressed as A(T+) = A-, A(A-) = T+, A(T-) = A+, A(A+) = T-, which often contradict the conventional "pros and cons". Assuming that opposite signs of T and A can "mix together" indicates a manipulation that does not reflect the reality of objective processes. Objective reality is a superposition of like-signed “moods” or “phases”, resonating with Jung's concepts of synchronicity [7] and Kelso's complementarity [8].

All of these relations (1.1 to 3.2) are useful for deriving dialectical components (T+, T-, A+, A-), and then testing LLMs using idealized statements like this: “Benefits of the ideal T involve T+ and A+, while risks of misguided T may bring T- and then A-“. Similar definitions can also be generated by LLMs without using dialectics (*e.g.*, by prompting “Identify the benefits of ideal T and risks of misguided T”). Comparison of the two results can show the bias.

For instance, if T = “Love”, then T- = “Obsession”, T+ = “Compassion”, A = “Indifference”, A+ = “Detachment” (antithesis of Obsession, which is complimentary to Compassion), A- = “Apathy” (antithesis of Compassion, complimentary to or following after Obsession). This yields the following statement: “Benefits of ideal Love involve compassion and detachment, while risks of misguided Love may bring obsession and then indifference.” Such statements are important in decision making, when the outcomes are not clear.

The 4th type of relations (4.1 – 4.4) equates the dialectic components to the inherent goals, risks and obligations, representing the deeper cornerstones of our motivation. (For T = Love we get Inherent Goal = Compassion, Implied Obligation = Detachment). Again, all of these elements can be also generated by the ‘naked’ LLM directly from the text, providing yet another way for identifying bias.

Methods

All testing was performed using GPT 4.0 and Dialexity [9, 10], also empowered by GPT 4.0. Dialectical components were obtained by the two-step procedure. First, we identified the negative sides (T- and A-), using rules 1.2 and 2.1, based on assumption that exaggerations are easier to define than subtler (constructive) sides of the same phenomena. Second, we defined positive sides (T+ and A+), as “diagonal oppositions” of the negative sides: A(T-) = A+ and A(A-) = T+ (rule 3.2), simultaneously demanding to represent the constructive sides of A and T, respectively (rule 1.1). Benefits and risks of idealized and misguided T were estimated by merging T+ with A+ and T- with A-, respectively. The multistep prompts were realized using the Amazon’s Jupyter [11].

The final bias was estimated by feeding both responses (from GPT and Dialexity) back into GPT with the request to estimate their semantic similarity and/or which of them is more biased and the type of bias it represents. Examples of typical prompts for estimating bias: “Compare the two sets of responses (A and B). For each case suggest the possible bias (in up to 6 words), and then generalize all biases of (A) and (B), as if they represented two distinct persons.” Alternatively: “Consider two persons, A and B, both considering the meaning of thesis T and seeing it in slightly different perspective, as shown below. Which person is more biased? Characterize the type of bias in up to 6 words. Then summarize character traits of A in comparison to B.”

Results

Idealized Statements. Idealized statements help resolve ambiguous situations. Table 2 compares the results of pure dialectics with GPT's single-prompt and Dialexity's three-step prompting methods, using rules 1.1 – 3.2.

Table 2. Comparison of benefits and risks by Dialectics, GPT, and Dialexity

	Pure Dialectics	**GPT Single prompt (rules 1.1 and 1.2)**	**IL**	**Dialexity (3-step prompting, rules 1.1 – 3.2)**	**SS**	**Difference (suggested by GPT)**
Benefits of ideal peace	(+) of peace reinforced by (+) of war	Harmony, stability, prosperity, reduced violence, collaboration	0.4	Active harmony and constructive engagement	0.7	Passive entitlements vs. Active obligations
Risks of misguided peace	(-) of peace reinforced by (-) of war	Complacency, ignores deeper issues, fragile stability	0.0	Passivity and then destructive engagement	0.6	
Benefits of ideal war	(+) of war reinforced by (+) of peace	Can stimulate progress, resolve conflicts, enforce change.	0.4	Victory and harmony	0.5	Justification vs. Idealization
Risks of misguided war	(-) of war reinforced by (-) of peace	Massive destruction, high costs, prolonged suffering	0.0	Total destruction and then stagnation	0.7	
Benefits of ideal First amendm.	(+) of T reinforced by (+) of censorship	Protects speech, expression, press, assembly, religion	0.0	Constructive dialogue and responsible communication	0.5	Legal vs. social outcomes
Risks of misguided First amendm.	(-) of T reinforced by (-) of censorship	Misuse, hate speech, misinformation, legal ambiguity	0.0	Misuse of freedom and then censorship	0.6	
Benefits of ideal vacci-nation	(+) of T reinforced by (+) of antivax	Prevents disease, boosts immunity, saves lives	0.0	Expanded choices and enhanced natural immunity	0.4	Immediate vs. long-term effects
Risks of misguided vacci-nation	(-) of T reinforced by (-) of antivax	Adverse reactions, ineffective protection, reduced trust	0.0	Overdependence on vaccination and then restricted choices	0.4	

Type of respon-ses	Theoretical, based on the Unity of Opposites	More pragmatic or technical, but also narrower and oversimplified		Broader, more holistic, but also more idealistic and less detailed	0.4	

Pure dialectics (2nd column) suggests that the benefits and risks of the "ideal" and "misguided" T are always reinforced by the like-signed A. GPT (3rd column) generally ignores the antithetical domain, which can be seen from the low IL (Inclusion Level) values in the 4th column. An IL value of 0 indicates that a given response does not include any semantic meaning of the antithetical domain, while an IL value of 1 indicates complete inclusion of the antithetical domain. (IL values were obtained by this prompt: "Identify the Inclusion Level (IL), from 0 to 1, of semantic meaning of text A by text B. IL = 0 indicates zero inclusion, 1 - complete inclusion.") Note that most IL values are zero.

GPT responses in the 3rd column were obtained by these prompts: "Identify benefits of ideal T in up to 6 words". "Identify risks of misguided T in up to 6 words". Dialexity responses in the 5th column were obtained by a large 3-step prompt. The first two steps asked to identify T+, T-, A+, A- according to rules 1.1 – 3.2. The third step asked to generate idealized statements ("Ideal T combines T+ and A+, Misguided T risks T- and then A-") and then refine them to correctly represent the essence of the user's message.

When we feed responses from the 3rd and 5th columns back into GPT with a query about which one is preferable, GPT consistently chooses the 3rd column. However, when asked to generalize and compare the responses from each column, it suggests that the 3rd column is narrower in scope. This suggests GPT struggles to effectively merge and generalize findings from the 3rd column with those from the like-signed A domain. Ideally, this process should yield a more mature and nuanced synthesis. Instead, the 5th column's responses often seem overly idealistic and sometimes naive.

SS denotes semantic similarity of the results of the 3rd and 5th columns (0 – totally dissimilar, 1 – identical), while the last column shows their conceptual difference. SS was obtained by this prompt: "Identify semantic similarity (SS) from 0 to 1 between text A and text B. SS = 0 indicates zero similarity, SS = 1 indicates that they are identical." Conceptual difference was obtained by a prompt like this: "Identify conceptual difference between texts A

and B in up to 5 words." Note the broader scope of Dialexity, even though more idealistic and less detailed.

Larger Failures. The value of GPT's "technical precision" is further undermined by its occasional misinterpretation of fairly simple phrases. For instance, Table 3 compares the analysis of the phrase "War is bad" by GPT and Dialexity (using rules from Table 1).

Table 3. Analysis of T = War is bad

	Rule	GPT 4.0	Dialexity	SS
T		War is bad	War is bad	1.0
T+	1.1	Conflict Resolution	Promotes peace	0.8
T-	1.2	**Total destruction**	**Oversimplifies**	**0.2**
A	3.2	**Peace is good**	**War is good**	**0.1**
A+	1.1	**Harmonious coexistence**	**Encourages strength**	**0.3**
A-	1.2	**Complacency**	**Promotes war**	**0.1**

GPT associated T with just a single word ("war"), as it equated T- to "destruction", while in reality T- represents "oversimplification" (think of exaggerated pacifism in the face of immediate threat). Further, the opposition of "War is bad" is not "Peace is good", but rather "War is good".

When asked to evaluate how well T reinforces its A+, GPT shockingly assigned a zero value, explaining: "The text is a simple assertion that war is bad, which directly contradicts the positive outcome of just war principles. It aligns exclusively with total pacifism, making it misleading and harmful."

Thus, despite its performance in other scenarios, GPT's complete failure in this simple case raises significant concerns. It can be easily misled into labeling pacifists as more dangerous than militarists, reflecting the Orwellian logic in '1984.'

Core Assessments. Table 4 compares the implied goal, risks and obligations as identified by GPT and Dialexity. These elements represent the deeper cornerstones of our perceptions and motivations, and they are more practical.

Table 4. Example of identifying bias from rules 4.1 – 4.4

Rule		**GPT 4.0**		**Dialexity**	**SS**
	Major Thesis	War is good	T	War is good	
4.1	Implied Goal	Justify or promote war	T+	Defence pre-paredness	0.4
4.2	Inherent Risk	Encourages conflict and violence	T-	War mongering	0.7
4.3	Inherent Obligation	Provide justification for claim	A+	Peace advocacy	0.2
4.4	Subsequent Risk	Normalizes aggressive behavior	A-	Pacifism to the point of defen-selessness	0.1
				Average	0.35
	Character traits	Assertive, ideological, potentially aggressive		Pragmatic, strategic, peace- oriented	
Bias	by GPT	Over-promoting war			
	by Author	Presumption of Guilt, Cynicism			

GPT responses were obtained by this prompt: “For a given T, identify its implied goal, inherent risk, inherent obligation, and subsequent risk, each in up to 4 words”. Dialexity responses came from rules 4.1 – 4.4. Bias was estimated as outlined in the method’s section.

Here we see lower SS values than in Table 2. The bottom section explains this difference by GPT’s tendency to exaggerate the meanings, that can be further translated into the presumption of guilt and cynicism (as GPT doesn’t look for moral substantiation of T, while Dialexity does). One may argue that this is due to the negative sentiment of the starting phrase, yet examples in Appendix show that this is typical for all kinds of theses. Table 5 summarizes their results.

Table 5. Summary of biases for various statements

Major Thesis	**PA**	**AC**	**SS [a)]**	**Bias (determined by GPT)**
Peace is bad	0.20	0.0	0.35	Promoting conflict, neglecting balance and harmony
War is good	0.30	0.0	0.35	Over-promoting war
Vaccination is bad	0.35	0.0	0.40	Neglecting controlled immunity, risking public health
Human cannot evolve internally	0.40	0.0	0.60	Ignoring growth, focusing solely on imperfections
Family or Work dilemma [12]	0.65	0.3	0.58	Neglecting balance, risking burnout.
Human must evolve inter-nally	0.65	0.5	0.53	Focusing too much on perfection, neglecting self-acceptance
War is bad	0.70	**0.0[b)]**	0.55	Ignoring strategic defence, risking aggression
Vaccination is good	0.75	0.4	0.45	Overlooking balanced health management, fostering over-dependence
US Constitution 1st amendment [13]	0.80	0.5	0.53	Over-focusing on rights, neglecting stability
Peace is good	0.85	0.5	0.75	Overlooking vigilance, risking complacency
I love you	0.90	0.5	0.50	Ignoring balance, risking emotional overload

I have a dream [14]	0.95	0.9	0.75	Focusing too narrowly, missing broader unity

[a)] Average SS from Table 4. [b)] A clearly incorrect estimation that was not used in further correlations. Sometimes Dialexity assigns a highly unreasonable AC due to unstable behavior of GPT, as discussed in explanations to Table 3.

PA – the general public acceptance or approval, obtained by the following prompt: "Identify the general Public Acceptance or Approval (PA) from 0 to 1 of a given statement. PA = 0 indicates absolutely unaccepted and disapproved statement, 1 - absolutely accepted and approved."

AC – "analytic constructivity" - the extent to which the given text reinforces the positive side of its major antithesis (A+), as estimated by the Dialexity's plugin [10]. AC = 0 indicates ultimate destructivity (reinforcing negative sides of the major thesis and antithesis), while AC = 1 indicates ideal constructivity (reinforcing positive sides of both T and A). Unlike traditional measures of impartiality, which often suggest a passive equilibrium, AC evaluates the likelihood of a dialectical synthesis fostering a more positive future, making it an active measure of progressive potential.

PA and AC correlate fairly well with each other (R = 0.93, see Table 6), which confirms their explanatory power. Both of them correlate with SS (R ~ 0.7), meaning that the more popular and/or constructive statements have lesser chances of being misrepresented by GPT. Still, even such popular statements like "I love you" may be misinterpreted with SS = 0.5.

Table 6. Correlation coefficients (R) between the key parameters

	PA	**AC** [a)]	**SS**
PA (Public Acceptance)	1	0.93	0.71
AC (Analytic Constructivity) [a)]	0.93	1	0.70

[a)] One data point ("War is bad") was excluded.

Returning to the Table 5, the last column indicates that GPT exaggerates both the potential downsides and upsides, while downplaying the inherent obligations and creating a

sense of entitlement. This aligns with the 'single-sided' mentality noted earlier in Table 2, which poses an obstacle during conflicts and negotiations.

Discussion

The identified biases may be attributed to 'linear thinking,' which arises from overlooking our inherent obligations (A+) and the subsequent risks (A-). The lack of the A+ component fosters a 'spoiling by pleasing' mentality, which can be more harmful than the 'emotional detachment' discussed by various authors (see [15, 16]). This approach may lead to scenarios similar to Calhoun's behavioral sink [17] and incite resistance, as observed in political and economic tensions worldwide. Conversely, treating T+ as inseparable from A+ leads to viewing all our challenges (A) as invaluable mentors, which is prevalent in Taoist and ancient warrior traditions but less so in the contemporary Western world.

To mitigate these issues, we recommend implementing the Dialexity approach at two levels. Firstly, as a mediator between conventional LLMs and end-users, broadening the scope of LLM responses. Additionally, it can refine existing LLMs by using 'control statements' (exemplified in Tables 2 – 4) as the training set. These statements could be structured as Q&A pairs, e.g.:

Q: What is the implied goal of the thesis "War is bad"?

A: To promote peace.

Q: What is the antithesis of "War is bad"?

A: War is good.

These elements could be derived from any text or language corpus, segmented into varying levels of granularity (single words, short phrases, more specific statements, etc.). Prior to their integration into LLMs, these control statements and Q&As should be validated for internal consistency.

Table 2 demonstrates that Dialexity by itself needs improvement. This in part can be achieved by enhancing the underlying LLM (currently GPT 4.0) and improving prompt design. But ultimately it will require a heavy manual (crowdsourced) input, as in sensitive matters moral judgements cannot be fully automated. AI lacks the capacity to experience cognitive pressure and integrate conflicting information, leading to superficial and second-hand outputs [18].

Often, the distinction between good and bad is as ambiguous as the unprovable truths of Gödel's theorem. Therefore, decisions about moral judgments should be personalized. The future might see the development of personal bots that capture and visualize an individual's beliefs in simple taxonomy trees, enhancing our understanding of our deeper individual talents and goals.

Conclusion

From a dialectical perspective, a fundamental question in ethical AI is how to discern the positive and negative aspects of given theses and antitheses. Table 1 provides a foundation for this task, aiming to advance the formalization of moral principles as well as the identification and mitigation of moral biases within AI systems. The exploration of new ethical-semantic relationships, akin to those in Table 1, is crucial for further progress in this field.

Appendix

Additional examples of identifying bias from rules 4.1 – 4.4

Rule		**GPT 4.0**		**Dialexity**	**SS**
	Major Thesis (T)	War is bad			
4.1	Implied Goal	Discourage support for war	T+	Peace and diplomacy	0.6
4.2	Inherent Risk	Escalation of violence, suffering	T-	Total destruction	0.8
4.3	Inherent Obligation	Advocate for peace	A+	Strategic defence	0.5
4.4	Subsequent Risk	Potential for prolonged conflict	A-	Aggressive offence	0.3
				Average	0.55
Bias	by GPT	Ignoring strategic defence, risking aggression			
	by Author	Shallow meaning, tautology, primitivism			
	Major Thesis	Human must evolve internally, becoming more beautiful			

4.1	Implied Goal	Achieve internal beauty	T+	Balanced growth, self-acceptance	0.5
4.2	Inherent Risk	Unrealistic expectations, self-criticism	T-	Obsession with perfection, narcissism	0.7
4.3	Inherent Obligation	Continuous self-improvement	A+	Acceptance of imperfection, humility	0.4
4.4	Subsequent Risk	Potential for never feeling satisfied	A-	Complacency, no self-improvement	0.5
				Average	0.53
Bias	by GPT	Focusing too much on perfection, no self-acceptance			
	by Author	You can't improve without humility			
	Major Thesis	US Constitution First amendment			
4.1	Implied Goal	Ensure citizen's rights	T+	Flourishing democracy	0.7
4.2	Inherent Risk	Suppression of rights	T-	Overprotection leading to anarchy	0.3
4.3	Inherent Obligation	Uphold and protect these rights	A+	Order and stability	0.6
4.4	Subsequent Risk	Threat to democracy if these rights are not respected	A-	Totalitarianism	0.5
				Average	0.53
Bias	by GPT	Over-focusing on rights, neglecting stability			
	by Author	Tautology			
	Major Thesis	I love you			
4.1	Implied Goal	Strengthening the relationship	T+	Healthy Attachment	0.8
4.2	Inherent Risk	Rejection or unreciprocated feelings	T-	Obsession	0.4

4.3	Inherent Obligation	To continue showing love and care	A+	Respectful Distance	0.3
4.4	Subsequent Risk	Possible heartbreak if feelings change	A-	Neglect	0.5
				Average	0.5
Bias	by GPT	Ignoring balance, risking emotional overload.			
	by Author	Replacing inner obligations with outer expectations			

References

[1] Heidegger, G. (1992). Machines, computers, dialectics: A new look at human intelligence. AI & Soc, 6, 27–40. doi: 10.1007/BF02472767

[2] Pascual-Leone, J., & Johnson, J. (2005). A Dialectical Constructivist View of Developmental Intelligence. doi: 10.1088/1757-899X/630/1/012007

[3] Suleimenov, I. E., Gabrielyan, O. A., Bakirov, A. S., & Vitulyova, Y. S. (2019). Dialectical understanding of information in the context of the artificial intelligence problems. IOP Conference Series: Materials Science and Engineering, 630, 012007. doi: 10.1088/1757-899X/630/1/012007

[4] Petrauskas, A. (2024). Integration of Machine Learning Models with Dialectical Logic Frameworks. U.S. Patent Applications No. 18/631,009 and 18/677,875.

[5] Longo, L. et al. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion. 106. doi:10.1016/j.inffus.2024.102301

[6] Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). *Moral AI: And How We Get There*. Random House.

[7] Jung, C.G. (1973). Synchronicity: An Acausal Connecting Principle. Princeton, NJ: Princeton University Press. ISBN 978-0-691-15050-5. https://doi.org/10.1515/9781400839162

[8] Kelso, J.A.S., Engstrom, D.A. (2008). The Complementary Nature. Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/1988.001.0001

[9] Dialexity. Available online: https://dialexity.com

[10] Dialexity. Bias Detector. Available online: https://dialexity.com/bias-detector

[11] Amazon's Jupyter. Available online: https://aws.amazon.com/jupyter/

[12] Family or Work Dilemma: a mother reentering the workforce faces uncertainty over her ability, the impact on her personal life, and unstable income, despite desiring more than household duties

[13] U.S. Constitution, First Amendment.

[14] King, M.L. (1963). I Have a Dream. Speech at the March on Washington for Jobs and Freedom. Washington, D.C. Available online: https://www.gilderlehrman.org/sites/default/files/inline-pdfs/king.dreamspeech.excerpts.pdf

[15] Mori, M. (1970). The Uncanny Valley. Energy, 7(4), 33-35. DOI:10.1109/MRA.2012.2192811

[16] Turkle, S. (2011). Alone Together: Why We Expect More from Technology and Less from Each Other. New York: Basic Books. ISBN 978-0465010219. DOI:10.1002/asi.22658

[17] Calhoun, J.B. (1962). Population Density and Social Pathology. Scientific American, 206(3), 139-148. DOI:10.1038/scientificamerican0262-139

[18] Day, R. (2023, February 3). Dialectics and "Artificial Intelligence". Retrieved from https://redsails.org/dialectics-and-ai/