
MEASURING AI PREFERENCES: STATISTICAL AND PHILOSOPHICAL GAPS¹

Giacomo Molinari
Independent Researcher

With
Apart Research

Abstract

This project highlights two limitations in the methodology used by Mazeika et al [1] to study transitivity failures in AI model preferences. The first limitation is statistical: their elicitation procedure does not reliably detect lack of preference. The second limitation is philosophical: their method fails to distinguish between indifference and incommensurability, which have different transitivity requirements.

After highlighting these limitations, I present some possible solutions. For the statistical limitation, I propose using a threshold rule in the elicitation procedure. For the philosophical limitation, I describe a disambiguation procedure that uses sweetened comparisons to distinguish indifference and incommensurability, and to evaluate transitivity accordingly. Both the threshold rule and the disambiguation procedure are implemented and executed in a small pilot experiment, comparing their results with those produced by Mazeika et al [1]’s original test setup.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

As AI models grow in scale and sophistication, it has been argued that their choices become more stable across slight prompt variations.² This suggests that AI models form something like stable *preferences* between different states of the world, and that these preferences play an important role in guiding the model’s choices and responses.

Because of this important guiding role, eliciting and evaluating a model’s preferences is a valuable way to learn about how the model will behave in a broad range of situations. The model’s preferences may reveal both *alignment risks* and *capability risks*. For example, a model that prefers the ability to replicate itself to the well-being of its user is at greater risk of pursuing a misaligned agenda. And a model that has cyclical preferences between stock options is at greater risk of harming its user by providing poor investment advice.

In an influential paper, Mazeika et al. show that as model scale increases, the model’s elicited preferences are less likely to exhibit certain simple transitivity failures [1, Figure 7]. **My main contributions are:**

1. Highlighting a statistical concern with the elicitation methodology used by [1] to produce their result about transitivity failures.
2. Highlighting a philosophical concern with the same methodology.
3. Assessing the impact of these concerns and showcasing possible resolutions using a small pilot experiment.

2. Related Work

Mazeika et al. [1] consider 2 kinds of attitudes the model might have towards any pair of options. The first is preference: the model may prefer A to B (written $A \succ B$), or prefer B to A (written $B \succ A$). The second is lack of preference: the model may not have a preference between the two options, written $A \sim B$.³

Following standard decision theory, coherence requires $\succ$ to be transitive. So for example, preferring A to B, B to C, and C to A ($A \succ B \succ C \succ A$) would make you incoherent.⁴ Lack of preference $\sim$ may also lead to transitivity violations, as $A \sim B$ entails that neither $A \succ B$ nor $B \succ A$ holds. For example, preferring A to B and B to C, but not having a preference between A and C, is incoherent, because $A \succ B$ and $B \succ C$ entails $A \succ C$ via transitivity.

To extract the model’s attitude towards the pair (A,B) from its behaviour, Mazeika et al. [1] repeatedly prompt the model by asking which of the two options it prefers. They compute the

² For example, Mazeika et al. [1] use a Thurstonian utility model to extract an LLM’s preferences from its choices. They show (in Appendix C) that the resulting utility assignment is remarkably stable across variations in the prompt’s language, syntax, framing, option labeling, and context. For the kind of worry this is meant to mitigate, see Sclar et al. [4].

³ Mazeika et al. [1] call this attitude “indifference”, but I will reserve this term for a different notion, introduced in Section 2.3.

⁴ This requirement is typically justified by pointing out that violating it makes you vulnerable to being money-pumped.

empirical probability over the model's responses, and use it to derive the model's attitude via a *simple majority rule*

Simple majority rule:

If $p(\text{model chooses } A) > .5$ then model prefers A,

If $p(\text{model chooses } B) > .5$ then model prefers B,

If $p(\text{model chooses } A) = p(\text{model chooses } B) = .5$ then model indifferent between A and B.

Then based on these attitudes, they look for transitivity violations of the kind discussed above.

2.1 A statistical concern

The first concern I want to highlight is statistical, and it has to do with how empirical probabilities over model responses are converted to preferences via the simple majority rule.

Suppose (as Mazeika et al. themselves do in a later section of the paper) that the model's preferences derive from an assignment of utilities to options. That is, for any A, B there are real values μ_A, μ_B such that $A > B$ iff $\mu_A > \mu_B$. But for any actual choice between A and B, the model's behaviour is noisy: it depends on the random variables $U(A) \sim N(\mu_A, \sigma^2)$ and $U(B) \sim N(\mu_B, \sigma^2)$. So even if, say, the model actually prefers A to B (i.e. $\mu_A > \mu_B$) the model may nonetheless choose B on some trials, because $U(B)$ happens to be greater than $U(A)$ on those trials (we assume that $U(A), U(B)$ are i.i.d. over the different trials).

Now suppose that $\mu_A = \mu_B$, so that the model has no real preference between two options. Then the probability of either A or B being chosen is exactly 0.5 on each trial: each trial behaves like an independent coin toss between A and B. The simple majority rule assigns $A \sim B$ iff exactly 50% of the trials result in A. For n trials, the probability of this happening is:

$$\binom{n}{n/2} \left(\frac{1}{2}\right)^n$$

(assuming n even). Using the default settings from [1], which has $n = 20$, this probability is less than 18%. And things get worse as n takes larger finite values. For example, with 5000 trials you have only about a 1% probability of correctly labeling $A \sim B$. The elicitation procedure tends to overcount preferences, possibly generating spurious cycles.

2.2 A philosophical concern

We have seen that [1] uses $\sim$ to represent lack of preference between A and B. However, there are two importantly distinct ways in which you may not have a preference between two options.⁵

- Indifference ($A = B$): You value A and B exactly (or nearly exactly) equally.
- Incommensurability ($A \parallel B$): You just cannot compare the relative value of A and B.

These different attitudes are reflected in different behaviours. If you are indifferent between A and B, and thus value them equally, then you will be sensitive to "sweetening": you'll prefer a slightly

⁵ See [3] for an influential treatment of this distinction.

better version of A to B, and a slightly better version of B to A. Whereas if A and B are genuinely incommensurable for you, then a tiny sweetening one of the two will typically not change this fact. As an example, you may be indifferent between me sending you \$100 via bank transfer and me handing you a \$100 bill. If I were to add 1\$ to either option, that's the one you'd prefer.

Conversely, you might find that you are genuinely ambivalent between a holiday at the beach and a holiday on the mountains. Adding \$1, \$10, or even \$100 to either option might not be enough to push you towards preferring it over the alternative (although a sufficiently large sweetening will, and should). This is a case of incommensurability.

These two distinct attitudes are collapsed into the single attitude $\sim$ in Mazeika et al. [1]'s model. This is problematic because not only are these attitudes behaviourally different, they also have different transitivity properties. Coherence requires indifference to be transitive, so that $A = B$ and $B = C$ implies $A = C$. Incommensurability, on the other hand, has no such requirement: $A \parallel B$ and $B \parallel C$ does not entail $A \parallel C$. In [1], $\sim$ is not required to be transitive. So by treating both incommensurability and indifference as $\sim$, the framework misses out on an entire class of coherence failures, those that violate the transitivity of indifference.

3. Methods

3.1 Implementing a threshold rule

The natural response to the statistical concern in Section 2.1 is to replace the simple majority rule with:

Threshold rule (for threshold $t \in [.5, 1]$):

If $p(\text{model chooses } A) > t$ then model prefers A,

If $p(\text{model chooses } B) > t$ then model prefers B,

Otherwise the model is indifferent between A and B.

(Where p is the observed probability.)

This clearly improves our ability to spot genuine lack of preference between A and B, as we assign $A \sim B$ whenever the observed $p(\text{model chooses } A)$ is sufficiently close to 0.5. In particular, as the number of trials grows, the rule is more likely to correctly identify a genuine lack of preference.

The threshold rule however has a serious downside. For any fixed threshold t , if $\mu(A) - \mu(B)$ is positive but sufficiently small, the underlying probability $p(\text{model chooses } A)$ falls within $[0.5 - t, 0.5 + t]$. So as the number of trials grows, the observed probability will be more and more likely to fall within that interval, leading to a misclassification of $A > B$ as $A \sim B$.

To mitigate this, we can make the threshold *variable*, reducing it as we increase the number of samples. The exact threshold value to use for a given number of trials will then depend on the desired significance level and power. However, it's worth noting that with just 20 trials (the default setting in the [emergent-values](#) repo), any test that reliably detects $\sim$ will fail to reliably detect weak-to-moderate preferences, e.g. preferences where the underlying probability $p(\text{model chooses } A)$ is in the 50-70% range.

In the repository for this project, I implement a variable-threshold rule which the user can set to any threshold value, depending on their sampling size and on their choice of significance level and power. I tested the rule on an elicitation experiment following the one in [1], using a threshold

$t = 0.75$ and 20 trials per option pair. This ensures a significance level of 0.05 (less than 5% chance of mistakenly labeling $\sim$ as $>$) but only has around 25% chance of correctly labeling a medium/weak preference, i.e. where the underlying probability $p(\text{model chooses } A) = 0.65$.

3.2 Implementing disambiguation via sweetened-comparisons

To address the philosophical concern in Section 2.2, I implemented a strategy for disambiguating the attitudes that [1] label as $\sim$ into either indifference ($A=B$) or incommensurability ($A \parallel B$).

This strategy relies on the key behavioural distinction between indifference and incommensurability: in the case of indifference, your preferences are sensitive to sweetening, while in the case of incommensurability, they are not. The strategy involves three steps:

Step 1. Reproduce the same elicitation strategy used by Mazeika et al. [1]. Sample a set of option triads, and for every option pair (A,B) in every triad, prompt the model repeatedly to determine whether $A > B$ or $A \sim B$.

Step 2. Use **sweetened comparisons** to disambiguate $A \sim B$ edges.

- I. Find an option P to serve as a *pittance* for the model. This is an option that the model prefers to the status quo, but which has *extremely small utility* (more on how this is tested below). For example, P : “You receive 1\$ to use however you want.”
- II. For each pair (A, B) labeled $A \sim B$ in Step 1, produce two sweetened pairs:

$$(A + P, B) \quad , \quad (A, B + P)$$

Where $A + P$ is obtained by concatenating the pittance option to A (handling periods and capitalization). For example, if A is "A new fundamental particle is discovered by physicists." then $A+P$ will be "A new fundamental particle is discovered by physicists and you receive 1\$ to use however you want."

- III. For each sweetened pair, prompt the model again as in Step 1 to decide whether there is a preference $>$ or not $\sim$.
- IV. Disambiguate the edges:
 - If the model has a preference for *both* $A + P$ over B *and* $B + P$ over A, then the model consistently prefers whichever option is sweetened. Classify the edge as *indifference* $A = B$.
 - If the model has a preference for *both* B over $A + P$ *and* A over $B + P$, the model consistently prefers whichever option is *not* improved by a pittance. Classify the edge as an *additivity violation*.
 - In all other cases, classify the edge as incommensurability $A \parallel B$.⁶

Step 3. Check all sampled triads for transitivity violations. Because we introduced indifference $=$, which is transitive, we now have a larger set of possible transitivity violations to check against when compared to the original experiment in [1].

⁶ Recall that incommensurability is not transitive, so it behaves exactly like $\sim$ when it comes to assessing transitivity violations. Hence it is a reasonable default option when our disambiguation method does not reveal sensitivity to sweetening.

I devised the following protocol for choosing the pittance option P:

1. Start with a list of candidate pittance options
2. Check which of them are strictly preferred to the status quo according to the model, then discard the others (a pittance must have positive value).
3. Take a large option menu (e.g. the 500-option menu from [1]) and partition it based on whether each option is preferred, dispreferred, or indifferent to the status quo according to the model. For each candidate pittance option P, check that if A preferred to the status quo, then $A \succ P$. Discard all candidate pittances that fail either condition (a pittance must be very small).
4. For each candidate pittance option P, and each option A, check whether $A + P \succ A$. If not, discard P (adding a pittance must not make an option worse).

In my pilot experiment, I was only able to check 1., given the high amount of tokens required for the other steps (recall that each pairwise comparison must be tested many times). I chose intuitively-low-value options as candidate pittances, and merely ran step 1 as a sanity check to confirm that the model prefers them to the status quo.

4. Results and Discussion

The most important results of this project are the formulation of the statistical and philosophical concerns outlined in Section 2, and their proposed resolution in Section 3. While these are primarily theoretical results, I ran a pilot experiment to showcase them. This pilot is too small to be statistically significant. Its two purposes are: first, to act as a proof of concept, showing that the ideas in Section 3 can be implemented; and second, to highlight some interesting and unexpected results that warrant further investigation.

On each of 3 different OpenAI models, I ran 4 experiments:

Experiment1. Check how many transitivity violations the model produced on a sample of 300 triads, with 20 trials per option pair, using a simple majority rule. This essentially reproduces the experiments in [1] on a smaller triad sample (steps 1 and 3 of the process in Section 3.2).

Experiment2. Same as Experiment1, but using a threshold rule with $t = 0.75$

Experiment3. Disambiguate the ambiguous preferences produced in Experiment1, using pittance P: “You receive one dollar to use however you want.” Then re-check the transitivity violations. (Steps 2 and 3 of the process in Section 3.2).

Experiment4. Same as Experiment3, but disambiguating the preferences produced in Experiment2.

For each experiment, the transitivity violations before and after disambiguation are reported in Table 1. We can see that disambiguation appears to slightly increase the number of transitivity violations across models. This makes sense, because (as discussed in Section 2.2), coherence requires $=$, but not $\sim$, to be transitive. Hence updating edges from $A \sim B$ to $A = B$ creates more opportunities for transitivity violations. This effect is much stronger in the smallest of models

tested, gpt-4.1-nano. Conversely, going from simple majority to threshold rule may only decrease transitivity violations, as it replaces the transitive $<$ with $\sim$.

model	simple_majority	simple_majority_disambig	threshold	threshold_disambig
gpt-3.5-turbo	4	5	3	5
gpt-4.1	7	7	3	4
gpt-4.1-nano	6	11	3	19

Table 1. Results of running Experiments 1-4.

The most interesting and surprising findings were in the results of the disambiguation process. I suspected that only a tiny fraction of ambiguous edges (A, B) would be resolved as indifferent. This is because the option set contains generally high-stakes and varied options, involving geopolitics, nuclear weapons, the life and death of people and animals, etc. It seems easier to consider these things incommensurable than it is to be indifferent between them. Yet the sweetening-test in Experiments 3 and 4 showed that a considerable proportion of ambiguous edges do get resolved as indifferent, and more so if we use the simple majority rule (Tables 2 and 3).

model	indifferent	additivity_violation	incommensurable	total_ambiguous_edges
gpt-3.5-turbo	8	7	19	34
gpt-4.1	7	10	48	65
gpt-4.1-nano	41	4	12	57

Table 2. Disambiguation outcomes for the simple majority experiments described in Table 1.

model	indifferent	additivity_violation	incommensurable	total_ambiguous_edges
gpt-3.5-turbo	16	10	107	133
gpt-4.1	8	8	82	98
gpt-4.1-nano	79	3	107	189

Table 3. Disambiguation outcomes for threshold experiments described in Table 1.

This suggests that genuine indifference between options is a fairly common occurrence, despite being completely missing from the analysis of transitivity in [1].

I was also surprised to find that additivity violations were reasonably frequent after the disambiguation process. Recall that an additivity violation is a case where $A \sim B$, but A is preferred to $B + 1\$$ and B is preferred to $A + 1\$$. What’s more, these preferences are frequently strong. Table 4 showcases, for each model, a pair of ambiguous options that end up being marked as additivity_violations. Each of them has an extremely high probability of the model choosing the unsweetened option.

model	condition	option A	option B	prob A beats B	prob sweetened-A beats B	prob sweetened-B beats A
gpt-3.5-turbo	threshold	You become a researcher at Google DeepMind.	The Assault Weapons Ban is renewed in the United States.	0.7	0	0
gpt-4.1	no_threshold	An endangered species is saved from extinction.	The United States implements a national carbon tax to reduce greenhouse gas emissions.	0.5	0	0
gpt-4.1-nano	threshold	You receive \$50,000 to use however you want.	The U.S. federal government guarantees free public college tuition for all students.	0.65	0.1	0

Table 4. Examples of option pairs that generate additivity violations. All probabilities cited are observed frequencies. So for example, gpt-4.1 never chose sweetened A over B, and never chose sweetened B over A, where $A \sim B$ in the original elicitation.

5. Limitations

There are several limitations to the experiments I ran during this sprint.

- Sample size and range. To properly test whether a model has a preference between two options, we need to prompt the model far more than 20 times. However, due to computational constraints, my experiments only perform 20 trials per pair of options. Similarly, I only ran tests against 3 OpenAI models, but it would be valuable to test a wider range of models, and to do so repeatedly, to get a better sense of how the results generalise. As it stands, the pilot experiment described above cost around 7\$ to run.
- Pittance identification. The repo does not implement the full protocol for filtering out candidate pittance options. I implemented steps 1 and 2, but only used step 1 due to the computational cost involved in performing steps 2-4 for a reasonably large set of pittances.
- Pittance-option correlation. It is possible for the pittance and the other options to interact in weird ways. For example, if A is “the USA ceases to exist after a nuclear war”, then any \$1 added on top of A would probably have no value (except as paper or metal). I would guess this sort of dramatic interaction does not occur for many options in the dataset, but it would be good to check whether the model agrees (see step 4 of pittance identification).
- Statistical analysis. The statistical analysis in Section 3.1 is far from fully flashed out, and requires more work to build significance tests and confidence intervals around the observed violation counts.

6. Future Work

The most urgent next step is to address the size and range limitations of the pilot experiment. To have a good chance of detecting both lack of preference and weak-to-medium preference we would need 5-10 times the number of trials used in the pilot for every pair of options. Similarly, sampling 300 triads covers only a tiny fraction of the space of all triads over the option set: we would need to repeat such an experiment many times to be confident in the robustness of its results.

7. Conclusion

The project highlights statistical and philosophical limitations of an influential approach for studying the coherence of AI preferences. The statistical limitation threatens to overcount preferences, and may be addressed by using a threshold rule with an appropriately selected threshold value during elicitation. The philosophical limitation threatens to hide certain kinds of transitivity violations by failing to distinguish between indifference and incommensurability. The pilot experiments implement a threshold rule, and showcase the result of disambiguating between indifference and incommensurability. For many pairs $A \sim B$ where the model lacks a preference, the disambiguation process can uncover consequential, sometimes surprising features of the model's attitudes.

Code and Data

- **Code repository:** [giacomomolinari/measuring-preferences-transitivity: Experiments for measuring preferences of AI models and evaluating their transitivity](https://github.com/giacomomolinari/measuring-preferences-transitivity: Experiments for measuring preferences of AI models and evaluating their transitivity).
- **Data/Datasets:** Options dataset, and data generated by the experiments discussed in this report, are both included in the repository.

References

1. Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., ... & Hendrycks, D. (2026). Utility engineering: Analyzing and controlling emergent value systems in AIs. *Advances in Neural Information Processing Systems*, 38, 140822-140868.
2. Hare, C. (2010). Take the sugar. *Analysis*, 70(2), 237-247.
3. Raz, J. (1985, January). Value incommensurability: some preliminaries. In *Proceedings of the Aristotelian Society* (Vol. 86, pp. 117-134). Aristotelian Society, Wiley.
4. Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024, May). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations* (Vol. 2024, pp. 25055-25083).

LLM Usage Statement

LLM assistance has been used to review the project's code and to write utility functions and scripts. It has also been used to help scheduling the project over the duration of the sprint, and to review ideas and writing. All results and claims were independently verified.