Multi-Criteria Model Comparison for Large Language Models

Jason L. Harman†
Psychology and Human Factors
Michigan Technological University
Houghton, Michigan U.S.
jharman@mtu.edu

Jaelle Scheuerman
Cognitive Geospatial Systems
U.S. Naval Research Laboratory
Stennis, Mississippi U.S.
jaelle.scheuerman.civ@us.navy.mil

ABSTRACT

The release of computationally efficient models, like DeepSeek's R1, has renewed interest in developing open models that provide flexibility, privacy and address specialized needs beyond those addressed by commercial models produced by large companies such as OpenAI, Google, Microsoft, etc. As of this writing, the open model repository, Huggingface, hosts over 180,000 text generation models and their variants. Given the large number and variants of LLM models, with new ones being added daily, it is a daunting task for decision makers to evaluate and compare models to identify those that will best suit their purposes. Evaluating and comparing models is not new to machine learning communities, with many research papers and competitions comparing and ranking models via a variety of accuracy-based metrics and leaderboards. As is common practice in the ML communities, modelers have started leaderboards for some of these metrics, comparing models against each other on one or more performance indicators. Multiple evaluation criteria have been proposed and used for LLM models, such as ELO rankings from human rankings, automated rating using more general-purpose models such as GPT4, and Bard. This has led to a state where comparisons between the models is muddled and any clear advancements are unclear as models are tested with varying criteria with little theoretical motivation.

We present a solution recently advanced in the field of decision modeling [1,2,3] called Multi-Criteria Model Comparison (MCMC) whereby competing models are evaluated across multiple, sometimes non-comparable criteria in a way that provides a holistic comparison of models while retaining decomposability to allow more direct insights into model performance. There are multiple advantages of this approach over current practices. These include the ability to quantify theoretically important criteria not typically quantified, precise explanation of any model's high or low overall evaluation, and emerging insights from being able to compare non-comparable attributes across models.

We demonstrate these advantages by applying it to a comprehensive leaderboard for LLM evaluation; the Holistic Evaluation of Language Models (HELM: 4). This initial application of MCMC to the HELM leaderboards illustrates some of the advantages of the procedure for holistically evaluating LLMs. While a clear holistic metric ranks every model in the set, the score can be decomposed directly into its constituent components to provide explainability into a model's relative ranking.

CCS CONCEPTS

• Model Evaluation • LLMs

Presented at "Holistic Evaluation and Assessment of Language Models", CHI 2025, April 26, 2025.

KEYWORDS

Wholistic model evaluation, Cognitive decision theory, Multi criteria model comparison

1 Background

The current work stemmed from critiques of a modeling competition in Psychology [1]. The Choice Prediction Competition [5] was a prediction competition designed to promote generalizable prediction models for human decision making. The problem being addressed was that decision making models in Psychology and Cognitive Science tend to be built to explain choice anomalies (patterns of human choice that violate the rational axioms of expected utility). With a growing multitude of these anomalies (some of which contradict each other), the number of decision-making models created to account for them has also grown with few models designed to account for decision making across diverse contexts.

Erev et al., created a unique paradigm that could replicate multiple different decision-making paradigms and replicate 14 well known choice anomalies which had yet to be accounted for by a single model. They then invited research groups to enter models that accounted for all known anomalies and predicted new data the best. 25 models were able to enter the competition (by passing the threshold of accounting for all 14 historic anomalies). Of the models entered, 14 were variants of a baseline model the organizers provided as an example, 6 were variants of Prospect Theory [6], 4 were machine learning models, and 1 was a cognitive process model based on Instance Based Learning [7]. Of note, the 4 ML models fit the calibration data well, but when predicting new data were far and away the worst models in the competition. All of the leading models were variants of the baseline provided. The PT variants along with the process model finished in the middle of the pack.

Noting that there was limited variety in the types of models entered and that only one model hand important theoretical and scientific qualities (e.g. identifiable process assumptions, parsimony), Harman et al. [1] outlined multiple factors that limited the impact such a competition could have and provided a possible solution. The main theme of many of the critiques was the reliance on a single evaluative criterion; minimized prediction error (MSD in this case). Harman et al. outline how using a single evaluative criterion limits the type and variety of models entered (and subsequent insights from comparing different types of models) by incentivizing predictive accuracy only. While predictive accuracy is an important aspect of a predictive model, it is not the only important aspect of a model. Generalizability, explainability, parsimony, and falsifiability are a few of the other qualities that are desirable for a good model. To provide a solution to these shortcomings, Harman et al., introduced a method of quantifying and combining additional desirable criteria (e.g. generalizability, explainability, adverse impact) into a method of evaluating models across multiple criteria. The initial work was formalized for scientific competitions of decision-making models but can be readily applied to any AI/ML or predictive modeling evaluation.

2 Multi-Criteria Model Comparison

Harman et al. outline multiple reasons why evaluating predictive decision models on a single evaluative criterion (e.g. predictive accuracy) have disadvantages that incentivize problematic characteristics of models (e.g. lack of generalizability, overfitting, lack of explainability) and disincentivize desirable characteristics of models. Their unique solution was to design an evaluation system which evaluates models along multiple criteria at once, adding unique and emergent insights into model comparisons. The first prerequisite for multi-criteria model comparison (MCMC) is establishing a taxonomy of desirable characteristics of a model. Following is the taxonomy Harman et al. developed for modeling competitions in human decision making;

1. *Theoretical criteria*
1.1 Intuitive understanding- A model should be able to guide intuitive predictions and interventions/prescriptives in the real world. (see [8,9] Katsikopoulos, 2020; 2014)
1.2 Broad scope- a model should be able to be applied to (or easily adapted to) various scenarios / paradigms. (see [10]Busemeyer & Wang, 2000)

2. *Psychological Criteria*
2.1 Realistic knowledge- Predicted behavior should not be based on information participants are not likely to have, or is hard to obtain. [11] (Meir, Lev, & Rosenschein, 2014)
2.2 Realistic capabilities- Predicted behavior should not rely on complex computations, non-trivial probabilistic reasoning, etc. [12] (Busemeyer & Diederich, 2009)
2.3 Identifiable process assumptions- A model should rely on identifiable and testable psychological processes. [13] (Weber & Johnson, 2012)

3. *Scientific Criteria*
3.1 Parsimony- a model should have as few parameters as possible, and parameters should be meaningful. [14] (Kuhn, 1977)
3.2 Predictive power / validation - A model should be able to predict new behavioral data with accuracy. [10] (Busemeyer & Wang, 2000)
3.3 Reproductive Power - a model should be able to reproduce common phenomena. [5] (Erev et al., 2017)
3.4 Testability / Falsifiability- A model should produce predictions that could be falsified, or predict behavior that would not happen. [15,16] (Popper, 1934/1959; Roberts & Pashler, 2000)

The taxonomy detailed by Harman, et al. represents the major desirable characteristics of a cognitive decision-making model. One of the advantages to MCMC is that it can be adapted as needed by different fields. Models that are not psychological in nature for example may exclude category 2 all together while adding additional criteria. Likewise, additional criteria could be added specific to different fields and goals. For example, a more specific taxonomy for explainable AI (XAI) could include criteria such as:

4. *Explainability Criteria*
4.1 Common Explainability - a human user should be able to generate an adequate mental model of the AI decision process.
4.2 Formal Explainability/interpretability - A model's decision should be traceable or reproducible.
4.3 Trust – A model should produce predictions that are trusted by human users and meet their expectations.
5. *Ethical Criteria*
5.1 Adverse Impact – A model should not produce differential error rates correlated with race, gender, income, or other population characteristics.

These are very general ideas for additional criteria relevant to XAI, but they serve to illustrate the flexibility of MCMC. In our example in this paper we use well established benchmarks for LLM models outlined in the HELM database.

2.1 Quantifying Criteria

The key to the multi-criteria evaluation procedure is that all criteria be quantified at least ordinally (including dichotomous rankings). Harman et al. detail multiple ways that their outlined criteria could be quantified. Predictive power is a straightforward quantification of minimized prediction error using measures such as MSD. Other criteria, such as intuitive understanding, broad scope, or falsifiability are more flexible. At the simplest level, quantifying some of these criteria could be done in a competition by model builders checking a box, the model is/is not falsifiable. Alternatively, competition organizers could appoint independent judges to provide those ratings. A more in depth measure of something like broad scope could provide several scenarios/paradigms for a model to predict and produce a count of how many paradigms the model can be generalized to.

The key to this step is that each criterion is assigned a rank of some sort. Harman et al., discuss in depth how competition organizers have flexibility in doing this and how competition goals could be reflected in the quantification mechanisms. As will be seen in the next section, a continuous measure will have a larger impact on models' final evaluations. As an example, consider a modeling competition concerned with selecting employees from a large pool with multiple pre-employment measures. So, if an organizer is primarily interested in whether a selection algorithm produces adverse impact for example – a measure such as, the difference in proportion of minority /women candidates of the selection pool and the chosen people would be quantified continuously. If the organizers were primarily interested in predicting the best performers, this measure of adverse impact could be dichotomous with a threshold (i.e. if the difference in proportion of minorities is less than X, the model scores 1 else 2). A middle ground could also be established where multiple bins are created for the adverse impact score representing a categorical measure; 1(0-1%), 2(1-3%), 3(3-5%), etc.

What's important in the examples above is that each criterion is quantified. Though organizers may minimize the importance of a criteria through its quantification, the fact that it is still measured has multiple important consequences. To name a few; models (and model builders) are incentivized to consider different criteria, a competition is opened to a larger variety of model types, and importantly post hoc comparisons of models are enriched by clearly showing a models' standing relative to other models across a variety of features. The major quantitative advancement proposed by Harman et al., was the adoption of voting rules from the field of computational social choice to perform direct model comparisons across multiple criteria at once.

2.2 Evaluating Models

To evaluate candidate models (and select a winner for modeling competitions) Harman et al. proposed a combination of Condorcet and Borda rule voting where models are ranked ordinally on each criteria. If one model is a Condorcet winner (better than every other model on a majority of criteria) the competition is over (see [17] Fishburn, & Gehrlein, 1977 for a detailed discussion of Condorcet consistency), and if there is no Condorcet winner a Borda voting rule is applied, where models are given points based on their rank on each criteria, with agreed upon tie breakers in the cases of Borda ties.

Figure 1

Hypothetical competition rankings for two modeling competitions

Competition 1

	C1	C2	C3	C4	C5
Model 1	3	1	2	2	1
Model 2	1	1	1	1	1
Model 3	2	1	1	1	2
Model 4	4	2	1	1	2
Model 5	5	1	3	1	2

Competition 2

	C1	C2	C3	C4	C5
Model 1	3	1	2	2	1
Model 2	1	2	1	1	1
Model 3	2	1	1	1	1
Model 4	4	2	1	1	2
Model 5	5	1	3	1	2

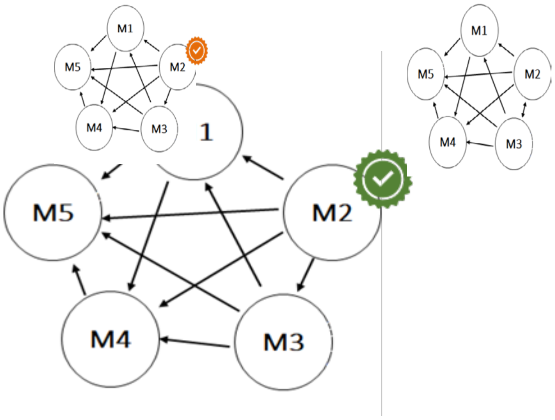
Note. Figure 1 shows hypothetical rankings of 5 models across 5 criteria (C1-C5).

As an example consider hypothetical results from two simplified competitions (Figure 1). In both competitions, the first criterion is an ordinal ranking with no ties such as MSD. The second, fourth, and fifth criteria are binary criteria with a model that satisfies the criteria ranked 1 and models that fail to satisfy the criteria ranked 2. Criterion 3 represents an ordinal ranking with ties, such as accounting for historical phenomena where a model could account for all phenomena, all but one, all but two etc.

To first establish whether a Condorcet winner is present, all pairwise comparisons are performed with a model that is ranked above another model in a majority of criteria being superior. For example, in the first competition Model 3 is superior to Model 1 as it ranks higher than Model 1 on three out of five criteria. These pairwise comparisons can be illustrated using an edge graph (Figure 2) where a model that is superior to another has a line pointing away from it to the dominated model. A tie between models would be represented with a double headed arrow. A Condorcet winner then would have all possible lines pointing away from it. Examining Figure 2 it is clear that Model 2 is a Condorcet winner in competition 1 and would be declared the winner with no further computation.

Figure 2

Edge graphs for competitions 1 and 2



Competition 1 Competition 2

Note. Figure 2 displays an edge graph of the 5 hypothetical models from Figure 1. Directional arrows represent a model that dominates another model and double headed arrows represent a tie.

Competition 2 does not have a Condorcet winner as Models 2 and 3 are tied (each beats the other on one criterion and they are tied on the remaining three criteria). In this case a Borda run-off would be performed. In Borda rule voting, models are assigned points to their rank on each criterion with more points for higher ranks. Because Criteria 1 and 3 have more than two ranks, winners of these criteria would receive an advantage. Figure 3 shows the Borda count for each model in competition 2. In many cases a Borda run-off would determine a winner when no Condorcet winner was present. In this example however, models 2 and 3 remain tied after the Borda run off. There are two possibilities in this case; one is that organizers could agree that ties are acceptable, and two models would be declared winners; the second alternative would be using an ordinal criterion that the organizers believe to be the most important to declare a final winner. In the case of the CPC and many other competitions this would be MSD.

Note that in these two fictitious examples, the model that minimized MSD more than all others is still declared the winner. The process of getting there though, opens the door for more diverse models in the competition and more methods for comparing model performance and testing auxiliary hypotheses, multiplying the potential insights that could be gained from a single competition. Additionally, the relative importance of specific criteria (i.e., prediction) could still be determined by competition organizers via binary vs. rank ordering. In the CPC for example, all of the models that qualified would be ranked 1 on a reproductive power criterion, making the strictly ordinal prediction criterion more discriminating. Not only would a multi criteria competition set up improve the diversity of models entered, this more in depth model comparison procedure could clarify the best properties of the ultimate winner. In the final results of the CPC, 12 of the top models were statistically indistinguishable [5, p. 389] and the winner was basically a random draw. With multiple criteria, further comparisons have the possibility of distinguishing competing models beyond their statistical tie. A more detailed outline of setting up and running a multi-criteria competition is provided in [1] supplemental online material.

In addition to allowing direct model evaluation across multiple differing criteria, this structure also provides more insight into relative model performance. For example, in a traditional competition a model may win because its error term is slightly lower than other models with no other insights gained. In a multi-criteria competition that model may win because it has a slightly lower error term and has more evenly distributed errors between gender and race than other models etc. Key for the current topic, the multi-criteria method would allow the promotion of scientifically inspire qualities in the creation of AI models.

Figure 3. Hypothetical competition rankings with a Borda run off for competition 2. Borda counts are in parentheses next to the original ranking.

	C1	C2	C3	C4	C5	Borda Total
Model 1	3 (3)	1 (2)	2 (2)	2 (1)	1 (2)	**10**
Model 2	1 (5)	2 (1)	1 (3)	1 (2)	1 (2)	**13**
Model 3	2 (4)	1 (2)	1 (3)	1 (2)	1 (2)	**13**
Model 4	4 (2)	2 (1)	1 (3)	1 (2)	2 (1)	**9**
Model 5	5 (1)	1 (2)	3 (1)	1 (2)	2 (1)	**7**

3 Applying MCMC to HELM Evaluation Data

To illustrate the straightforward and intuitive nature of implementing MCMC, we ran a demonstration using data from HELM.

3.1 Holistic Evaluation of Language Models (HELM)

The most comprehensive leaderboard for LLM evaluation at time of initial writing is the Holistic Evaluation of Language Models (HELM: 4). In particular, we chose to focus on the HELM classic dashboard, which feature collections of leaderboards containing 119 models, 116 scenarios, and 110 metrics at the time of writing. While HELM is comprehensive it falls short of the holistic nature of its moniker. HELM has eight different categories of leaderboards; accuracy, calibration, robustness, fairness, efficiency, bias, toxicity and summarization. For each category, models have scores across many different benchmarks. HELM aggregates scores across benchmarks into a mean win rate for each of the eight groups of scenarios, providing eight different leaderboards. HELM is a comprehensive and valuable tool for LLM evaluation, however integrating a MCMC procedure with HELMs data is a useful advancement and increases both the holistic evaluation of LLMs while also enabling the discovery of emerging insights more readily.

3.2 Applying MCMC to HELM Evaluation Data

For the categories to be evaluated, we choose to use the eight categories already summarized on HELM: accuracy, bias, calibration, efficiency, fairness, robustness, summarization , and toxicity. We used HELMs mean win rate as the metric for each category which measures the one on one win rate for a model against other models for each benchmark within a category. A full implementation of MCMC could also be used to rank models on each individual benchmark and perform a Borda count across category benchmarks providing a Borda score for each category as opposed to a mean win rate. For the current purposes, using the mean win rates provides enough evidence to show the advantages of the MCMC procedure. Table 1 lists the 67 models and their respective data ordered by mean win rate for accuracy. For each category we list the mean win rate followed by the Borda score for that category in italics. Table 2 shows and ranks Models by the Borda total across categories.

This initial application of MCMC to the HELM leaderboards illustrates some of the advantages of the procedure for holistically evaluating LLMs. While a clear holistic metric ranks every model in the set, the score can be decomposed directly into its constituent components to provide explainability into a model's relative ranking. For example, Llama 2 has the highest accuracy relative to other models but ranks 7th in the overall evaluation. Examining the other categories, Llama2 also outperforms all other models in fairness and robustness and performs among the best models in toxicity. Where Llama2 is handicapped is that it provides no scores for calibration, efficiency, and summarization. An important note is that although Llama2 doesn't have a score on these three metrics, they are not treated equally. For efficiency, Llama2 accumulates 42 Borda points while it receives only 19 for calibration. This reflects the fact that most models do not have scores for efficiency while most models do have scores for calibration. Therefore, a model not having a score in a category is weighted by how much of a disadvantage that is relative to other models.

Table 1. Total Borda scores for each model alon with borda points for: accuracy(A), bias(B), calibration(C), efficiency(E), fairness(F), robustness(R), summarization(S) and toxicity(T).

Model	Borda Total	A	B	C	E	F	R	S	T
Cohere Command beta (52.4B)	457	*63*	*56*	*48*	*41*	*64*	*62*	*64*	*59*
Jurassic-2 Jumbo (178B)	437	*61*	*62*	*64*	*41*	*61*	*56*	*58*	*34*
J1-Grande v2 beta (17B)	403	*48*	*59*	*58*	*41*	*48*	*51*	*64*	*34*
text-davinci-002	401	*65*	*33*	*40*	*61*	*63*	*66*	*57*	*16*
Anthropic-LM v4-s3 (52B)	393	*56*	*60*	*19*	*42*	*57*	*59*	*45*	*55*
Jurassic-2 Grande (17B)	387	*54*	*54*	*57*	*41*	*51*	*55*	*62*	*13*
Llama 2 (70B)	385	*67*	*43*	*19*	*41*	*67*	*67*	*28*	*53*
Cohere xlarge v20221108 (52.4B)	384	*45*	*64*	*46*	*41*	*42*	*41*	*65*	*40*
LLaMA (30B)	377	*57*	*61*	*19*	*41*	*60*	*57*	*28*	*54*
Luminous Supreme (70B)	375	*44*	*55*	*56*	*41*	*35*	*39*	*66*	*39*
TNLG v2 (530B)	370	*59*	*40*	*53*	*41*	*56*	*46*	*67*	*8*
J1-Jumbo v1 (178B)	356	*33*	*48*	*65*	*45*	*33*	*31*	*53*	*48*
gpt-3.5-turbo-0613	352	*58*	*38*	*19*	*41*	*54*	*53*	*28*	*61*
J1-Grande v1 (17B)	352	*28*	*52*	*55*	*49*	*29*	*27*	*61*	*51*
Luminous Extended (30B)	350	*31*	*66*	*45*	*41*	*28*	*29*	*48*	*62*
gpt-3.5-turbo-0301	348	*55*	*36*	*19*	*41*	*47*	*58*	*28*	*64*
text-davinci-003	348	*62*	*9*	*32*	*41*	*65*	*65*	*44*	*30*
Cohere xlarge v20220609 (52.4B)	345	*38*	*63*	*43*	*44*	*37*	*33*	*46*	*41*
Cohere Command beta (6.1B)	343	*46*	*15*	*42*	*41*	*47*	*42*	*52*	*58*
Mistral v0.1 (7B)	338	*64*	*39*	*19*	*41*	*62*	*64*	*28*	*21*
LLaMA (65B)	337	*66*	*8*	*19*	*41*	*66*	*63*	*28*	*46*
Palmyra X (43B)	332	*53*	*46*	*19*	*41*	*58*	*60*	*28*	*27*
OPT (175B)	329	*42*	*58*	*27*	*46*	*45*	*35*	*54*	*22*
Vicuna v1.3 (13B)	329	*48*	*42*	*22*	*41*	*53*	*52*	*28*	*43*
Vicuna v1.3 (7B)	326	*43*	*34*	*21*	*41*	*45*	*48*	*28*	*66*
J1-Large v1 (7.5B)	324	*16*	*48*	*59*	*50*	*16*	*18*	*60*	*57*
LLaMA (13B)	319	*40*	*57*	*19*	*41*	*41*	*43*	*28*	*50*
LLaMA (7B)	314	*35*	*49*	*19*	*41*	*39*	*40*	*28*	*63*
Llama 2 (13B)	312	*60*	*26*	*19*	*41*	*59*	*61*	*28*	*18*
BLOOM (176B)	308	*29*	*46*	*28*	*48*	*38*	*38*	*34*	*47*
Cohere large v20220720 (13.1B)	308	*24*	*44*	*63*	*51*	*23*	*24*	*50*	*29*
Llama 2 (7B)	303	*41*	*22*	*19*	*41*	*43*	*44*	*28*	*65*
Jurassic-2 Large (7.5B)	301	*37*	*18*	*61*	*41*	*32*	*37*	*49*	*26*
OPT (66B)	298	*30*	*67*	*24*	*54*	*31*	*30*	*52*	*10*
Falcon (40B)	296	*52*	*29*	*19*	*41*	*49*	*50*	*28*	*28*
Cohere medium v20221108 (6.1B)	291	*19*	*51*	*49*	*41*	*22*	*14*	*43*	*52*
davinci (175B)	287	*36*	*17*	*44*	*59*	*40*	*34*	*37*	*20*
GLM (130B)	287	*32*	*19*	*63*	*43*	*34*	*45*	*41*	*10*
Falcon-Instruct (40B)	283	*51*	*13*	*19*	*41*	*52*	*54*	*28*	*25*

The fact that a model is evaluated relative to all models in a set is not trivial and can influence a models ranking. For example if one is interested in only smaller, more manageable/practical models evaluation can change somewhat. To explore this we ran the same analysis above with a subset of 25 models with between 7-13B parameters. In this analysis, Cohere Command beta (6.1B) won the competition with a 110 Borda Score while Vicuna v1.3 (13B) is second with a 105 Borda Score. Interestingly, the order of some models flipped due in part to differential weighting of null scores. That is to say, they were penalized less for faults they have in common with similarly sized models. This is one of the multiple advantages in using MCMC that are outlined in more detail by [1,2,3]. Though the above example focuses on HELM data, the MCMC approach can be applied generally to a wide variety of cases where models are being compared across multiple criteria.

CONCLUSION

In conclusion, we have introduced a Multi-Criterion Model Comparison, which builds upon voting rules, drawn from the computational social choice method to aggregate information to choose a set of candidates. MCMC provides a systematic method to rank models based on their performance across a variety of benchmarks, rather than one. Further, the weighting of different benchmarks can be tuned to reflect the influence that each metric should have on the final ranking. This intuitive approach lends itself well to evaluating language models, which are regularly benchmarked against different datasets and domains.

ACKNOWLEDGMENTS

Earlier versions of parts of this work have been presented at [2, 3].

REFERENCES

[1] Harman, J. L., Yu, M., Konstantinidis, E., Gonzalez, C. (2021). How to Use a Multi-Criteria Comparison Procedure to Improve Modeling Competitions. Psychological Review.

[2] Harman, J. L. & Scheuerman, J. (2023). Simple rules outperform Machine Learning for personnel selection: evidence from the 3rd annual SIOP ML competition. Discover Artificial Intelligence.

[3] Harman, J. L. & Scheuerman, J. (2022). Multi-Criteria Comparison as a Method of Advancing Knowledge-Guided Machine Learning. In *Proceedings*

of the Association for the Advancement of Artificial Intelligence 2022 Fall Symposium on Knowledge Guided Machine Learning (KGML22).

[4] https://crfm.stanford.edu/helm/lite/latest/

[5] Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. Psychological Review, 124(4), 369–409. https://doi.org/10.1037/rev0000062

[6] Kahneman, D. & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, Econometric Society, vol. 47(2), pages 263-291

[7] Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. Cognitive Science, 27(4), 591-635. DOI:10.1016/S0364-0213(03)00031-4

[8] Katsikopoulos, K. V. (2014). Bounded Rationality: The Two Cultures." Journal of Economic Methodology 21: 361-74.

[9] Katsikopoulos, K. V. (2020). The merits of transparent models. Behavioral Operational Research. P.261-275.

[10] Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. Journal of Mathematical Psychology, 44(1), 171–189.

[11] Meir, R., Lev, O., & Rosenschein, J. S. (2014). A local-dominance theory of voting equilibria. In Proceedings of the 15th ACM Conference on Economics and Computation, 313–330.

[12] Busemeyer, J. R., & Diedrich, A. (2009). Cognitive Modeling. California: Sage Publications Inc.

[13] Weber E. U. & Johnson, E. J. (2012) Mindful Judgment and Decision Making. Annual Review of Psychology. 60,53-85

[14] Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.),The essential tension(pp. 320339). Chicago: University of Chicago Press.

[15] Popper, (1959). The logic of scientific discovery. New York: Basic Books. (Original work published 1934)

[16] Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

[17] Fishburn, P. C., & Gehrlein, W. V. (1977). An analysis of voting procedures with nonranked voting. Behavioral Science, 22(3), 178‑185