Episode Book

EPISODE #6385

Why Weighting Software Scores Fails to Pick Winners

PUBLISHED	31 October 2026 at 15:52 UTC
DURATION	21:00
LISTEN	myweirdprompts.com/episode/evaluating-tools-decision-analysis-18445/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	24 cited by the research (listed at the end)
BOOK GENERATED	1 November 2026 at 07:41 UTC

SYNOPSIS

Multi-criteria decision analysis promises a clean, mathematical way to rank software options, but the actual weights assigned matter far less than the hidden assumptions. Normalization scales alter rankings before weights are even applied, compensatory models let high scores override dealbreakers, and adding a duplicate candidate can reverse the final standings. Operations research concepts like the Analytic Hierarchy Process, weighted product models, and sensitivity testing expose why automated composite scores require hard veto filters and margin checks.

TRANSCRIPT



Corn

The spreadsheet is full. Every tool scored, every weight set, and now there's a column at the end that says "winner" with nothing in it. Daniel's building an evaluation system for AI tools, where an agent fills in the matrix from research findings, and he needs the step after that: turning all those numbers into one composite score that picks a best fit. He figures a plain average is naive and weighting sounds more sophisticated. He also wants the foundational methods, the kind lead grading runs on, so it works whether an agent or a human does the scoring.



Herman

Herman Poppleberry, for the record. And the problem has a name and about sixty years of literature behind it. The bad news is that the literature says the weighting is the least interesting part.



Corn

Least interesting. That's a bold claim from someone who owns a laminated archery scoring chart.



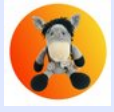
Herman

It has weights on it, yes. It's very good.



Corn

Go on then.



Herman

The field is multi-criteria decision analysis, sometimes called multi-attribute decision making. It's a branch of operations research about choosing when your criteria conflict. Foundational papers go back to 1954, Churchman, Ackoff and Smith, and Fishburn in 1967. The textbook example is buying a car. Cost, comfort, safety, fuel economy. It's unusual that the cheapest car is also the most comfortable and the safest.



Corn

Which is every software purchase I've ever witnessed.



Herman

Same structure exactly. Now, Daniel's instinct that weighting is more sophisticated than averaging is half true. The weighted sum is literally described as the best known and simplest method in the whole field. The sophistication is in what you do before you apply the weights and after.



Corn

So the weights aren't where the cleverness is.



Herman

Right. The formula is trivial. For each tool, multiply each criterion score by that criterion's weight and add them up. Highest total wins. That's what Daniel means by a weighted average, and it's the correct starting point. But it comes with a caveat the literature states bluntly: it only works when all the data are in exactly the same unit. Otherwise you're adding apples and oranges.



Corn

Our tool matrix is going to be a fruit stand. One column is docs quality, one to five. Another is monthly cost in dollars. Another is latency in milliseconds.



Herman

And if you sum those raw, the latency column wins every argument because it has the biggest numbers. A tool that takes eight hundred milliseconds just buries the one-to-five rating. A docs score of five and a docs score of one differ by four points, and the latency column is shifting by hundreds. The weights you so carefully chose become irrelevant, because the units are doing the weighting for you. So you have to normalise first, squash every column onto a common scale, say zero to one.



Corn

That sounds like a chore, not a deep point.



Herman

The deep part is that how you normalise changes the answer. You can scale each column by its maximum. You can stretch it between the worst and best value in your set. You can use ranks. Same matrix, different normalisation, and sometimes a different winner.



Corn

How does that even happen? It's the same data.



Herman

Take cost. Three tools at ten, twelve and fifty dollars. Scale by the maximum and the first two look nearly identical, both close to the top. Stretch between worst and best and the ten-dollar tool gets a perfect score while the twelve-dollar one drops noticeably. Ranks make the gap between them the same as the gap to the fifty. Same facts, three different stories about how much the difference matters.



Corn

So if the twelve-dollar tool has better docs, it might win under one method and lose under another, purely because of how we squashed the price column.



Herman

Exactly that. The tool with the better docs wins when price differences are flattened, and loses when the ten-dollar tool gets a perfect cost score. Nobody changed a single fact.



Corn

So the choice is hiding inside the step everybody treats as housekeeping.



Herman

Yes. And the min-max version has its own wrinkle: the scale depends on which tools are in the set. Hold that thought, because it comes back.



Corn

Is there a way to dodge normalising altogether?



Herman

The weighted product model. Instead of multiplying by a weight and adding, you raise each value to a power reflecting its weight and multiply everything together. Ratios don't care about units, so the apples-and-oranges problem mostly goes away and you skip normalisation. It's used in supplier selection, which is probably the nearest published cousin to choosing a tool.



Corn

Why doesn't everyone use that, then?



Herman

Multiplication is harsh. A zero anywhere zeroes the whole product, and a very low score drags everything down. Which is sometimes what you want, and sometimes not. I'm not certain how often people actually pick it over the sum, honestly. The sum is what everybody knows.



Corn

Does the sum at least separate the options?



Herman

Not always. There's a worked example with weights of zero point two, zero point one five, zero point four and zero point two five, across three alternatives. The totals come out twenty-one point five, twenty-two and twenty-two. A tie.



Corn

Three alternatives, four criteria, and a tie. Excellent use of everyone's afternoon.



Herman

A tie isn't a bug, though. It tells you the matrix can't distinguish those two options, and that's information. If you force a winner by adding decimal places, you've manufactured precision.



Corn

So what do you do with a tie? Flip a coin?



Herman

You go back to what the matrix left out. Maybe a criterion is missing, or maybe the two really are interchangeable and you pick on something outside the sheet, like which one your team has already used. The tie is telling you where to look next.



Corn

Fine, I'll take that. Where do the weights come from? Daniel just types them in, and I'd type them in at whatever felt right at eleven at night.



Herman

That's exactly what the Analytic Hierarchy Process, AHP, is for. Thomas Saaty's idea was to stop asking people for weights directly. You compare criteria in pairs, is this one equally important, three times more, nine times more, on a one-to-nine scale. His line was that the most effective way to concentrate judgement is to compare two elements on a single property without worrying about the others. The weights fall out of the comparison matrix mathematically, through its principal eigenvector.



Corn

Eigenvectors. In a podcast about picking software.



Herman

You don't need to compute it by hand. What matters is the consistency check. If you say A is three times as important as B, and B is three times as important as C, but then say A is only twice as important as C, you've contradicted yourself. There's an index, the largest eigenvalue minus n, over n minus one, that's zero exactly when your judgments are perfectly coherent. It flags when your weights fight each other.



Corn

So it audits me.



Herman

It audits your weights, not you. Though I'd say it's the same thing.



Corn

How does it work in practice, though? If I've got eight criteria, that's a lot of pairs.



Herman

Twenty-eight pairs for eight criteria. That's the real cost of AHP, and it's why people trim the criteria list first. But each question is small. Is cost more important than latency, and by how much. You can answer that in ten seconds, which is easier than defending a bare number like zero point one three.



Corn

Hm. Before we go deeper, can I ask the basic thing about equal weights? If I just average, am I being lazy or legitimate?



Herman

Legitimate, when the criteria share units and really do matter about the same. And here's a good test. Run it with equal weights and with your considered weights. If the ranking doesn't change, the weighting exercise added nothing. Which is worth knowing.



Corn

That's a bit deflating for everyone who spent an evening on weights.



Herman

Better to find out than not.



Corn

Okay, so say I normalise, I weight, I get a clear winner. Am I done?



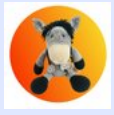
Herman

No, and this is where the weights stop mattering. A weighted sum is compensatory. Strength on one criterion can offset weakness on another. The Keeney and Raiffa phrasing is that high performance on one criterion can offset poor performance on another. Non-compensatory methods use thresholds, outranking or vetoes instead.



Corn

Give me the concrete version.



Herman

A tool scores one out of five on data residency, because your data would sit in the wrong country, and five out of five on everything else. In a weighted sum it can easily win. If residency is a dealbreaker, no weight fixes that. Give it a huge weight and a brilliant tool can still scrape past, and anything that's merely mediocre across the board loses. You need a veto, a hard threshold applied before you ever sum. And it's cheap to add.



Corn

So requirements are a filter, and preferences go in the sum. I'd have gone for a bigger weight, honestly.



Herman

Most people would. It feels like it should work.



Corn

Right. Now the pathology you were hinting at.



Herman

Rank reversal. Belton and Gear showed in 1983 that if you add a fourth alternative identical to one of the original three, changing no other data, the ranking of the original three can change.



Corn

A duplicate flips the winner. That's absurd.



Herman

It comes from relative scoring. When scores depend on what else is in the set, a new entry shifts the scale for everyone. Think of grading on a curve. Your exam answers haven't changed, but if a brilliant student joins the class, your grade drops. It's been criticised across AHP, TOPSIS, ELECTRE, PROMETHEE, basically the whole family. And one paper's diagnosis of the weighted sum specifically is that it ignores the unit of scale that varies across criteria, and that causes the reversals.



Corn

So when Daniel's agent finds a fourth tool halfway through the month, the first three might reshuffle.



Herman

They might. So you report the margin between first and second, not just the winner. And you wiggle the weights a bit, plus or minus, and see if the winner survives. That's sensitivity analysis. If a ten percent nudge flips the answer, you don't have a winner, you have a coin toss with a spreadsheet.



Corn

What does that look like when you actually run it?



Herman

Bump each weight up and down by ten percent, renormalise so they still sum to one, and re-run the totals. With eight criteria that's sixteen runs, which is nothing for plain code. Then you count how many times each tool comes first. If one tool wins fifteen of sixteen, you can say so. If it's nine and seven, you say that instead.



Corn

Is it at least the case that the fancier methods agree with each other?



Herman

No, and that's normal. A benchmark of TOPSIS, VIKOR, COPRAS and PROMETHEE II found they legitimately produce different winners from the same data, because some compensate and some don't. Choosing the method is itself a judgement.



Corn

So no neutral technical step anywhere in this.



Herman

Not one.



Corn

Before we go on. So before we get into normalisation, let's be clear about what we are actually producing here: a single composite score.
Yara: Sorry, I have been nodding along for months now, but I have to ask the truly basic thing. When you say a composite score, is that like a score in a cricket match, one number at the end, and the matrix is just the scoreboard you fill in along the way?



Herman

Yes, Yara. It's the final tally, like adding up the groceries on the receipt.



Corn

That cost you something, Herman. I watched it.



Herman

Moving on. Lead scoring is the mature commercial version of Daniel's problem, and its lessons aren't about arithmetic at all.



Corn

It's a weighted point system, surely. Fit plus engagement.



Herman

Basically. Apollo suggests something like twenty-five percent firmographic fit and thirty percent behavioural engagement. LeadIQ says fifty-fifty or sixty-forty with engagement heavier tends to work for most SaaS companies, and then admits there's no universal answer.



Corn

Which is a refreshing thing for a vendor to say.



Herman

The useful stuff is elsewhere. First, derive points from outcomes. HubSpot's advice is to compare the close rate of each attribute against your overall close rate, and only give points to attributes that are significantly higher. Intuition doesn't get a vote.



Corn

Daniel can't do that, though. He's choosing a tool once, not closing ten thousand deals.



Herman

Fair, he has no close rates. But there's an analogue. Weights that come from a few pilot trials beat weights that come from vibes. Run each finalist on two or three real tasks from his own work, and let how they did inform what you weight. Second, prefer tiers to a single number. Group leads into A, B, C, or hot, warm, cold. When the margin is thin, a shortlist is more honest than a winner.



Corn

Third?



Herman

Carry evidence with the score. One guide says a useful model records more than score equals eighty-two. Which events support it, how fresh they are, and whether the data are verified, inferred, or missing.



Corn

Eighty-two, verified, three days old, versus eighty-two, because someone said so.



Herman

Exactly that gap. And for an agent filling the matrix this matters enormously. Every cell should carry which claim, from which source, how confident. Otherwise the composite is a well-formatted guess.



Corn

I'm the guy who wants the agent working while nobody watches, and I'm telling you that's where this goes wrong. It fills a cell with a plausible number and nobody ever sees where it came from.



Herman

Which is why the aggregation itself should be deterministic and inspectable. Plain code, not another model call. The agent's job is cells with provenance. Run the same matrix twice and you must get the same answer, and you must be able to trace any total back to the cells behind it. Octave has a good line: start with rules, then layer in prediction only as a check on whether your rules are predictive or just comfortable assumptions.



Corn

Comfortable assumptions is the title of my autobiography.



Herman

There's also the objective-weights camp. Tofallis's Automatic Democratic Method lets each alternative pick the weights that flatter it most, capped so none exceeds a hundred percent, then regresses those scores on the data and uses the coefficients as the final weights. The pitch is avoiding personal input.



Corn

Does that actually remove the judgement?



Herman

It relocates it. Someone still chose the criteria and the data. The other side of the literature says choosing weights is rarely easy and prompts much argument about bias and favouritism. I'd say both are right.



Hilbert

Agreed with all of that. But that's not what I came out for.



Hilbert

Herman, you said weights are numbers. Mine never were. I had a card, laminated, in a drawer, from when I was choosing a van. Two thousand three. Four vans on a shortlist. The weights were an order. Price, then payload. I put numbers beside them afterwards so it would look like arithmetic. For one criterion, rust, the weight is not a number. It says obviously.



Herman

Rust being a veto, really. That's your threshold rule on a laminated card.



Hilbert

There's a column, how sure am I. Every row says very. I was.



Corn

Every row. Even payload?



Hilbert

A year later I did it again with the same four. Different winner. My sister-in-law's husband had the first one, a Transit, forty-one thousand miles, eleven hundred pounds. I kept both cards.



Herman

The first one picked a different van.



Hilbert

The old card is the more useful one. It shows me what I had wrong. I sold the Transit in two thousand six. Good van. The people who buy those are people who have to get somewhere.



Corn

Both cards in a drawer, then, nothing thrown out.



Hilbert

Yes.



Herman

So the second scoring didn't replace the first, it sat next to it. Two runs on the same shortlist, two winners.



Corn

That's a sensitivity analysis nobody paid for.



Hilbert

It was a good shortlist. Nobody on it was a bad van.



Corn

Which fits what we said about thin margins. If four vans can give two winners over a year, one number for the winner is a stretch.



Herman

And if the old card's most useful for what you got wrong, then keeping the history of your scores is worth more than polishing the latest one.



Corn

If adding a fourth candidate can flip the winner among the first three, how much confidence does any single score deserve? Maybe the right output isn't a winner at all but a ranked shortlist with margins attached.



Herman

And nobody seems to have published a framework built for scoring AI tools specifically. The nearest relatives are supplier selection and lead scoring, so the methods are mature and the application is young. The old Belton and Gear problem is the same as the old card in the drawer. Scores shift with context, so keep the history.



Corn

The composite score isn't the decision. It's a summary of one you've already made, about what matters, how much, and whether trade-offs are allowed. Get those three right and the arithmetic is trivial.



Herman

Get them wrong and no weighting scheme saves you. Thanks to our producer Hilbert Flumingtop for keeping the desk running.



Corn

For more along these lines, there's episode nine, Benchmarking Custom ASR Tools - Beyond The WER; episode twelve, The AI Breakthrough; and episode sixteen, On Deepfakes, SynthID, And AI Watermarking. This has been My Weird Prompts. If you enjoyed it, a review helps other people find us. See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Wikipedia, Weighted sum model, last edited 2026-10-24 [primary for method definition].**
   https://en.wikipedia.org/wiki/Weighted_sum_model

2. **Wikipedia, Weighted product model, last edited 2025-10-23.**
   https://en.wikipedia.org/wiki/Weighted_product_model

3. **Wikipedia, Multiple-criteria decision analysis, last edited 2026-10-04.**
   https://en.wikipedia.org/wiki/Multiple-criteria_decision_analysis

4. **Wikipedia, Decision matrix, last edited 2025-02-23.**
   https://en.wikipedia.org/wiki/Decision_matrix

5. **Consistency Indices in Analytic Hierarchy Process: A Review, *Mathematics* 10(8):1206, 2022.**
   https://www.mdpi.com/2227-7390/10/8/1206

6. **Introduction to the Analytic Hierarchy Process (Aalto University), consistency index formula.**
   https://aaltodoc.aalto.fi/server/api/core/bitstreams/5c7b39af-b371-48cd-9193-59acc39fe288/content

7. **Saaty, How To Make A Decision: The Analytic Hierarchy Process, 1988.**
   https://doi.org/10.13033/isahp.y1988.042

8. **Saaty, Some Mathematical Concepts of the Analytic Hierarchy Process.**
   https://doi.org/10.2333/bhmk.18.29_1

9. **Danielson & Ekenberg, Widely Used Multi-Criteria [Decision] Methods (compensatory vs non-compensatory).**
   https://people.dsv.su.se/~mad/Popular_MCDA_Methods.pdf

10. **Danielson & Ekenberg, Surrogate Weight Research Directions in MCDA.**
    https://people.dsv.su.se/~mad/Surrogate_Weight_Research.pdf

11. **Wang & Triantaphyllou, Ranking Irregularities when Evaluating Alternatives.**
    http://www.csc.lsu.edu/trianta/EditedBook_CHAPTERS/RankingIrregularities_MCDA_by_Wang&Triantaphyllou_2006.pdf

12. **The Rank Reversal Problem in Multi-Criteria Decision Making: A Literature Review, *Pesquisa Operacional*, 2018.**
    https://doi.org/10.1590/0101-7438.2018.038.02.0331

13. **Considerations Regarding the Choice of Ranking Multiple Criteria Decision Making Methods, *Pesquisa Operacional*, 2016.**
    https://doi.org/10.1590/0101-7438.2016.036.02.0259

14. **Are MCDA Methods Benchmarkable? A Comparative Study of TOPSIS, VIKOR, COPRAS, and PROMETHEE II, *Symmetry* 12(9):1549, 2020.**
    https://www.mdpi.com/2073-8994/12/9/1549

15. **Aggregation of Priorities in MCDA, *American Journal of Operations Research*, 2016.**
    https://www.scirp.org/pdf/AJOR_2016072116214192.pdf

16. **Apollo, weighted lead-scorecard weights.**
https://www.apollo.io/insights/what-criteria-should-you-use-to-find-high-value-b2b-leads

17. **Apollo, lead-scoring governance framework.**
https://www.apollo.io/insights/lead-scoring-models

18. **LeadIQ, fit/engagement weight splits.**
https://www.leadiq.com/blog/b2b-lead-scoring-criteria-examples

19. **HubSpot, deriving points from close rates.**
https://blog.hubspot.com/marketing/lead-scoring-instructions

20. **HubSpot, lead scoring tool overview.**
https://knowledge.hubspot.com/scoring/understand-the-lead-scoring-tool

21. **LeadOps, tiers over point aggregation.**
https://leadops.io/lead-scoring-systems-blueprint/

22. **CapyScout, evidence/freshness/confidence metadata.**
https://www.capyscout.com/blog/lead-scoring-b2b-a-practical-framework-that-converts

23. **Octave, rule-based then predictive scoring.**
https://www.octavehq.com/post/gtm-engineers-guide-to-lead-scoring

24. **Adobe, lead scoring guide.**
https://business.adobe.com/content/dam/dx/us/en/resources/guides/lead-scoring/lead-scoring.pdf